APA Handbooks in Psychology

APA Handbook of Testing and Assessment in Psychology

Kurt F. Geisinger, *Editor-in-Chief*

SHOUT
SHARE
IT

APA Handbook of

Testing and Assessment in Psychology

APA Handbook of Testing and Assessment in Psychology, Vol. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology, edited by K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez

APA Handbooks in Psychology

APA Handbook of Testing and Assessment in Psychology

VOLUME 1

Test Theory and Testing and Assessment in Industrial and Organizational Psychology

Kurt F. Geisinger, *Editor-in-Chief*

Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodriguez, *Associate Editors*

American Psychological Association • Washington, DC

Copyright © 2013 by the American Psychological Association. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, including, but not limited to, the process of scanning and digitization, or stored in a database or retrieval system, without the prior written permission of the publisher.

Published by
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242
www.apa.org

To order
APA Order Department
P.O. Box 92984
Washington, DC 20090-2984
Tel: (800) 374-2721; Direct: (202) 336-5510
Fax: (202) 336-5502; TDD/TTY: (202) 336-6123
Online: www.apa.org/pubs/books/
E-mail: order@apa.org

In the U.K., Europe, Africa, and the Middle East, copies may be ordered from
American Psychological Association
3 Henrietta Street
Covent Garden, London
WC2E 8LU England

American Psychological Association Staff
Gary R. VandenBos, PhD, *Publisher*
Julia Frank-McNeil, *Senior Director, APA Books*
Theodore J. Baroody, *Director, Reference, APA Books*
Lisa T. Corry, *Project Editor, APA Books*

Typeset in Berkeley by Cenveo Publisher Services, Columbia, MD

Printer: United Book Press, Baltimore, MD
Cover Designer: Naylor Design, Washington, DC

Library of Congress Cataloging-in-Publication Data

APA handbook of testing and assessment in psychology / Kurt F. Geisinger, editor-in-chief ; Bruce A. Bracken . . . [et al.], associate editors.
v. cm. — (APA handbooks in psychology)
Includes bibliographical references and index.
Contents: v. 1. Test theory and testing and assessment in industrial and organizational psychology — v. 2. Testing and assessment in clinical and counseling psychology — v. 3. Testing and assessment in school psychology and education.
ISBN 978-1-4338-1227-9 — ISBN 1-4338-1227-4
1. Psychological tests. 2. Psychometrics. 3. Educational tests and measurements. I. Geisinger, Kurt F., 1951- II. Bracken, Bruce A. III. American Psychological Association. IV. Title: Handbook of testing and assessment in psychology.
BF176.A63 2013
150.28′7—dc23
2012025015

British Library Cataloguing-in-Publication Data
A CIP record is available from the British Library.

Printed in the United States of America
First Edition

DOI: 10.1037/14047-000

Contents

Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology

Editorial Board ix
About the Editor-in-Chief xi
Contributors xiii
Series Preface xxi
Introduction xxiii

Part I. Test Theory 1

Chapter 1. Psychometric Characteristics of Assessment Procedures: An Overview 3
Anita M. Hubley and Bruno D. Zumbo
Chapter 2. Reliability 21
Kurt F. Geisinger
Chapter 3. The Generalizability of Test Scores 43
Edward W. Wiley, Noreen M. Webb, and Richard J. Shavelson
Chapter 4. Test Validity 61
Stephen G. Sireci and Tia Sukin
Chapter 5. Factor Analysis of Tests and Items 85
Li Cai
Chapter 6. Applying Unidimensional Item Response Theory Models to Psychological Data 101
Steven P. Reise, Tyler M. Moore, and Mark G. Haviland
Chapter 7. Item Analysis 121
Randall D. Penfield
Chapter 8. Bias in Psychological Assessment and Other Measures 139
Jeanne A. Teresi and Richard N. Jones
Chapter 9. Test Development Strategies 165
Neal M. Kingston, Sylvia T. Scheuring, and Laura B. Kramer
Chapter 10. Item Banking, Test Development, and Test Delivery 185
David J. Weiss
Chapter 11. Scaling, Norming, and Equating 201
Michael J. Kolen and Amy B. Hendrickson

Chapter 12. Basic Issues in the Measurement of Change 223
John J. McArdle and John J. Prindle
Chapter 13. The Standards for Educational and Psychological Testing 245
Daniel R. Eignor
Chapter 14. Technical Reporting, Documentation, and the Evaluation of Tests 251
Jane Close Conoley, Collie W. Conoley, and Rafael Julio Corvera Hernandez
Chapter 15. Ethics in Psychological Testing and Assessment 265
Frederick T. L. Leong, Yong Sue Park, and Mark M. Leach
Chapter 16. The Importance of Editorial Reviews in Ensuring Item Quality 283
Cathy Wendler and Jeremy Burrus
Chapter 17. Fairness Review in Assessment 293
Michael J. Zieky

Part II. Types of Testing 303

Chapter 18. Objective Testing of Educational Achievement 305
Michael C. Rodriguez and Thomas M. Haladyna
Chapter 19. Objective Personality Testing 315
Samuel E. Krug
Chapter 20. Performance Assessment in Education 329
Suzanne Lane
Chapter 21. Language Testing: History, Validity, Policy 341
Tim McNamara

Part III. Industrial and Organizational Psychology 353

Chapter 22. Assessment in Industrial and Organizational Psychology: An Overview 355
John P. Campbell
Chapter 23. Work Analysis for Assessment 397
Juan I. Sanchez and Edward L. Levine
Chapter 24. Thinking at Work: Intelligence, Critical Thinking, Job Knowledge, and Reasoning 417
Nathan R. Kuncel and Adam S. Beatty
Chapter 25. Biographical Information 437
Neal Schmitt and Juliya Golubovich
Chapter 26. Assessment of Leadership 457
Nancy T. Tippins
Chapter 27. Understanding and Improving Employee Selection Interviews 479
Robert L. Dipboye and Stefanie K. Johnson
Chapter 28. Personality Measurement and Use in Industrial and Organizational Psychology 501
Leaetta M. Hough and Brian S. Connelly
Chapter 29. Work Sample Tests 533
George C. Thornton III and Uma Kedharnath
Chapter 30. Situational Judgment Measures 551
Robert E. Ployhart and Anna-Katherine Ward

Chapter 31. Holistic Assessment for Selection and Placement 565
Scott Highhouse and John A. Kostek
Chapter 32. Employment Testing and Assessment in Multinational Organizations 579
Eugene Burke, Carly Vaughan, and Ray Glennon
Chapter 33. Performance Appraisal 611
Kevin R. Murphy and Paige J. Deckert
Chapter 34. Implementing Organizational Surveys 629
Paul M. Connolly
Chapter 35. Counterproductive Work Behaviors: Concepts, Measurement, and Nomological Network 643
Deniz S. Ones and Stephan Dilchert
Chapter 36. Stereotype Threat in Workplace Assessments 661
Ann Marie Ryan and Paul R. Sackett
Chapter 37. Job Satisfaction and Other Job Attitudes 675
Reeshad S. Dalal and Marcus Credé
Chapter 38. Legal Issues in Industrial Testing and Assessment 693
Paul J. Hanges, Elizabeth D. Salmon, and Juliet R. Aiken

Editorial Board

EDITOR-IN-CHIEF

Kurt F. Geisinger, PhD, Director, Buros Center for Testing, W. C. Meierhenry Distinguished University Professor, Department of Educational Psychology, and Editor, *Applied Measurement in Education*, University of Nebraska–Lincoln

ASSOCIATE EDITORS

Bruce A. Bracken, PhD, Professor, School Psychology and Counselor Education, College of William and Mary, Williamsburg, VA

Janet F. Carlson, PhD, Associate Director and Research Professor, Buros Center for Testing, University of Nebraska–Lincoln

Jo-Ida C. Hansen, PhD, Professor, Department of Psychology, Director, Counseling Psychology Graduate Program, and Director, Center for Interest Measurement Research, University of Minnesota, Minneapolis

Nathan R. Kuncel, PhD, Marvin D. Dunnette Distinguished Professor, Department of Psychology, and Area Director, Industrial and Organizational Psychology Program, University of Minnesota, Minneapolis

Steven P. Reise, PhD, Professor, Chair of Quantitative Psychology, and Codirector, Advanced Quantitative Methods Training Program, University of California, Los Angeles

Michael C. Rodriguez, PhD, Associate Professor, Quantitative Methods in Education, Educational Psychology, and Director, Office of Research Consultation and Services, University of Minnesota, Minneapolis

About the Editor-in-Chief

Kurt F. Geisinger, PhD, is currently director of the Buros Center on Testing and W. C. Meierhenry Distinguished University Professor of Educational Psychology at the University of Nebraska–Lincoln. He has previously been professor and chair of the Department of Psychology at Fordham University; professor of psychology and dean at the State University of New York at Oswego; professor of psychology and academic vice president at LeMoyne College; and professor of psychology and vice president for academic affairs at the University of St. Thomas, Houston, TX. His primary interests lie in validity theory, admissions testing, proper test use, the use of tests with individuals with disabilities, the testing of language minorities, the translation or adaptation of tests from one language and culture to another, and outcomes assessment. He has been a board member at large for the American Psychological Association (APA); a representative of the Division of Evaluation, Measurement, and Statistics to the APA Council of Representatives; an APA delegate and chair of the Joint Committee on Testing Practices (1992–1996); a member of APA's Committee on Psychological Testing and Assessment; chair of the National Council on Measurement in Education's (NCME's) Professional Development and Training Committee; cochair of NCME's Program Committee (1994); chair of the Graduate Record Examination Board; chair of the Technical Advisory Committee for the Graduate Record Examination; a member of the SAT Advisory Committee; a member and chair of the College Board's Advisory Research Committee; a member of NCME's Ad Hoc Committee to Develop a Code of Ethical Standards Committee; and has served on numerous other ad hoc task forces and panels. He is a fellow of APA, the Association for Psychological Science, and the American Educational Research Association. He is currently editor of *Applied Measurement in Education* and is currently serving or has served on the editorial committees for the *International Journal of Testing, Educational and Psychological Measurement, College Board Review, Educational Measurement: Issues and Practice, Psychological Assessment, Practical Assessment: Research and Evaluation, Journal of Educational Research,* and *Improving College and University Teaching.* He has edited or coedited the *Psychological Testing of Hispanics*, *Test Interpretation and Diversity,* and *High Stakes Testing in Education: Science and Practice in K–12 Settings* and the 17th and 18th *Mental Measurements Yearbooks* and *Tests in Print*.

Contributors

Jamal Abedi, PhD, School of Education, University of California, Davis
Bashir Abu-Hamour, PhD, Department of Counseling and Special Education, Mu'tah University, Alkarak, Jordan
Maria Acevedo, PhD, Department of Psychological and Educational Services, Fordham University, New York, NY
Phillip L. Ackerman, PhD, School of Psychology, Georgia Institute of Technology, Atlanta, GA
Juliet R. Aiken, PhD, Georgetown Law Center, Georgetown University, Washington, DC
Craig A. Albers, PhD, Department of Educational Psychology, University of Wisconsin–Madison
Elizabeth M. Altmaier, PhD, Department of Psychological and Quantitative Foundations, University of Iowa, Iowa City
Christopher T. Barry, PhD, Department of Psychology, University of Southern Mississippi, Hattiesburg
Adam S. Beatty, Doctoral Candidate, Department of Psychology, University of Minnesota, Minneapolis
Courtney A. Bell, PhD, Educational Testing Service, Princeton, NJ
Margit I. Berman, PhD, Department of Psychiatry, Dartmouth Medical School, Lebanon, NH
Nancy E. Betz, PhD, Professor Emeritus, Department of Psychology, The Ohio State University, Columbus
Christopher P. Borreca, JD, Thompson & Horton LLP, Houston, TX
Elizabeth A. Borreca, EdD, School of Education, University of St. Thomas, Houston, TX
Bruce A. Bracken, PhD, School Psychology and Counselor Education, College of William and Mary, Williamsburg, VA
Derek C. Briggs, PhD, School of Education, University of Colorado at Boulder
Shawn Bubany, PhD, Counseling Center, State University of New York College at Oneonta
Eugene Burke, MSc, SHL Group Ltd., Surrey, United Kingdom
Matthew K. Burns, PhD, Educational Psychology, College of Education & Human Development, University of Minnesota, Minneapolis
Jeremy Burrus, PhD, Educational Testing Service, Princeton, NJ
James N. Butcher, PhD, Professor Emeritus, Department of Psychology, University of Minnesota, Minneapolis

Heather A. Butler, Doctoral Candidate, School of Behavioral & Organizational Sciences, Claremont Graduate University, Claremont, CA
Li Cai, PhD, Graduate School of Education and Information Studies and the Department of Psychology, University of California, Los Angeles
Wayne Camara, PhD, The College Board, New York, NY
Gregory Camilli, PhD, School of Education, University of Colorado at Boulder
John P. Campbell, PhD, Department of Psychology, University of Minnesota, Minneapolis
Cindy I. Carlson, PhD, Department of Educational Psychology, University of Texas at Austin
Janet F. Carlson, PhD, Buros Center for Testing, University of Nebraska–Lincoln
Jerry Carlson, PhD, Professor Emeritus, Graduate School of Education, University of California, Riverside
Gail M. Cheramie, PhD, School Psychology Program, University of Houston–Clear Lake
Ruth A. Childs, PhD, Department of Leadership, Higher and Adult Education, University of Toronto, Toronto, Ontario, Canada
Ting-Wei Chiu, PhD, School of Education, University of Colorado at Boulder
Brian S. Connelly, PhD, Department of Management, University of Toronto Scarborough, Toronto, Ontario, Canada
Paul M. Connolly, PhD, Performance Programs Inc., Old Saybrook, CT
Collie W. Conoley, PhD, Department of Counseling, Clinical, and School Psychology, Gevirtz Graduate School of Education, University of California, Santa Barbara
Jane Close Conoley, PhD, Department of Counseling, Clinical, and School Psychology, Gevirtz Graduate School of Education, University of California, Santa Barbara
Marcus Credé, PhD, Department of Organizational Sciences and Communication, The George Washington University, Washington, DC
Reeshad S. Dalal, PhD, Department of Psychology, George Mason University, Fairfax, VA
Matthew Daley, Doctoral Candidate, College of Education, University of Florida, Gainesville
Paige J. Deckert, Doctoral Candidate, Department of Psychology, The Pennsylvania State University, University Park
Bridget V. Dever, PhD, College of Education, Georgia State University, Atlanta, GA
Bryan J. Dik, PhD, Department of Psychology, Colorado State University, Fort Collins
Stephan Dilchert, PhD, Department of Management, Baruch College, The City University of New York, New York, NY
Robert L. Dipboye, PhD, Department of Psychology, University of Central Florida, Orlando, FL
Neil J. Dorans, PhD, Educational Testing Service, Princeton, NJ
Ron Dumont, EdD, NCSP, School of Psychology, Fairleigh Dickinson University, Teaneck, NJ
Tanya L. Eckert, PhD, Department of Psychology, Syracuse University, Syracuse, NY
Daniel R. Eignor, PhD, Educational Testing Service, Princeton, NJ
Kadriye Ercikan, PhD, Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Vancouver, British Columbia, Canada
Giselle B. Esquivel, PsyD, ABPP, Department of Psychological and Educational Services, Graduate School of Education, Fordham University, New York, NY
Stephen E. Finn, PhD, Center for Therapeutic Assessment, Austin, TX
Paul J. Frick, PhD, Department of Psychology, University of New Orleans, New Orleans, LA

Kurt F. Geisinger, PhD, Buros Center for Testing and Department of Educational Psychology, University of Nebraska–Lincoln

Drew H. Gitomer, PhD, Graduate School of Education, Rutgers, The State University of New Jersey

Ray Glennon, PhD, SHL Group Ltd., Surrey, United Kingdom

Juliya Golubovich, Doctoral Candidate, Department of Psychology, Michigan State University, East Lansing

William M. Grove, PhD, Department of Psychology, University of Minnesota, Minneapolis

Thomas M. Haladyna, PhD, Professor Emeritus, Arizona State University, Tempe

Diane F. Halpern, PhD, Department of Psychology, Claremont McKenna College, Claremont, CA

Ronald K. Hambleton, PhD, Department of Educational Policy, Research and Administration and Center for Educational Assessment, School of Education, University of Massachusetts, Amherst

Paul J. Hanges, PhD, Department of Psychology, University of Maryland, College Park

Jo-Ida C. Hansen, PhD, Department of Psychology, Counseling Psychology Graduate Program, and Center for Interest Measurement Research, University of Minnesota, Minneapolis

Virginia Smith Harvey, PhD, College of Education and Human Development, University of Massachusetts Boston

John Hattie, PhD, Melbourne Graduate School of Education, University of Melbourne, Carlton, Victoria, Australia

Beth E. Haverkamp, PhD, Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Vancouver, British Columbia, Canada

Mark G. Haviland, PhD, Department of Psychiatry, School of Medicine, Loma Linda University, Loma Linda, CA

Kirk Heilbrun, PhD, Department of Psychology, Drexel University, Philadelphia, PA

Amy B. Hendrickson, PhD, The College Board, Newtown, PA

Rafael Julio Corvera Hernandez, MA, Department of Counseling, Clinical, and School Psychology, Gevirtz Graduate School of Education, University of California, Santa Barbara

Bridget O. Hier, Doctoral Candidate, Department of Psychology, Syracuse University, Syracuse, NY

Scott Highhouse, PhD, Department of Psychology, Bowling Green State University, Bowling Green, OH

Jill S. Hill, PhD, Department of Counseling and Clinical Psychology, Teachers College, Columbia University, New York, NY

Stephanie Brooks Holliday, Doctoral Candidate, Department of Psychology, Drexel University, Philadelphia, PA

Sandra L. Horn, Doctoral Candidate, Department of Psychology, University of Toledo, Toledo, OH

Leaetta M. Hough, PhD, The Dunnette Group, Ltd., St. Paul, MN

Anita M. Hubley, PhD, Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Vancouver, British Columbia, Canada

Stephen S. Ilardi, PhD, Department of Psychology, University of Kansas, Lawrence

Susan Jacob, PhD, Department of Psychology, Central Michigan University, Mount Pleasant
Stefanie K. Johnson, PhD, College of Business, University of Colorado, Denver
Richard N. Jones, ScD, Institute for Aging Research at Hebrew SeniorLife; and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA
Randy W. Kamphaus, PhD, College of Education, Georgia State University, Atlanta, GA
Irvin R. Katz, PhD, Educational Testing Service, Princeton, NJ
Alan S. Kaufman, PhD, Yale Child Study Center, Yale University School of Medicine, New Haven, CT
Uma Kedharnath, Doctoral Candidate, Department of Psychology, Colorado State University, Fort Collins
H. Elizabeth King, PhD, Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA
Neal M. Kingston, PhD, Department of Psychology and Research in Education and Center for Educational Testing and Evaluation, University of Kansas, Lawrence
Jennifer L. Koehler, Doctoral Candidate, Department of Psychology, Syracuse University, Syracuse, NY
Elizabeth A. Koenig, Doctoral Candidate, Department of Psychology, Syracuse University, Syracuse, NY
Michael J. Kolen, PhD, Department of Psychological and Quantitative Foundations, The University of Iowa, Iowa City
Rebecca Kopriva, PhD, Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison
Kathleen B. Kortte, PhD, ABPP-CN/RP, Johns Hopkins Outpatient NeuroRehabilitation Program and Department of Physical Medicine and Rehabilitation, The Johns Hopkins University School of Medicine, Baltimore, MD
John A. Kostek, Doctoral Candidate, Department of Psychology, Bowling Green State University, Bowling Green, OH
Laura B. Kramer, PhD, Center for Educational Testing and Evaluation, School of Education, University of Kansas, Lawrence
Samuel E. Krug, PhD, MetriTech, Inc., Champaign, IL
Lauren S. Krumholz, PhD, Department of Psychology, Harvard University, Cambridge, MA
Nathan R. Kuncel, PhD, Department of Psychology and Industrial and Organizational Psychology Program, University of Minnesota, Minneapolis
Michael J. Lambert, PhD, Department of Psychology, Brigham Young University, Provo, UT
Suzanne Lane, PhD, Department of Psychology in Education, University of Pittsburgh, Pittsburgh, PA
Mark M. Leach, PhD, Department of Educational and Counseling Psychology, University of Louisville, Louisville, KY
Heidi Leeson, PhD, Monocle Consulting Ltd., Auckland, New Zealand
Frederick T. L. Leong, PhD, Department of Psychology, Consortium for Multicultural Psychology Research, Michigan State University, East Lansing
Melanie E. Leuty, PhD, Department of Psychology, University of Southern Mississippi, Hattiesburg
Edward L. Levine, PhD, Dipl., Professor Emeritus, Department of Psychology, University of South Florida, Tampa, FL

Pei-Ying Lin, PhD, Department of Educational Psychology and Special Education, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
Richard M. Luecht, PhD, Department of Educational Research Methodology, University of North Carolina at Greensboro
Juliette Lyons-Thomas, Doctoral Candidate, Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Vancouver, British Columbia, Canada
Sara Maltzman, PhD, County of San Diego Child Welfare Services, San Diego, CA
Kobus Maree, PhD, DEd, DPhil, Department of Educational Psychology, University of Pretoria, Groenkloof, Pretoria, South Africa
Hale Martin, PhD, Graduate School of Professional Psychology, University of Denver; and the Colorado Center for Therapeutic Assessment, Denver, CO
Mark E. Maruish, PhD, LP, Southcross Consulting, Burnsville, MN
Shawn N. Mason, MA, Department of Educational Psychology, University of Minnesota, Minneapolis
Nancy Mather, PhD, Department of Disability and Psychoeducational Studies, University of Arizona, Tucson
John J. McArdle, PhD, Department of Psychology, University of Southern California, Los Angeles
R. Steve McCallum, PhD, Department of Educational Psychology and Counseling, University of Tennessee, Knoxville
Carina McCormick, Doctoral Candidate, Educational Psychology Department, University of Nebraska–Lincoln
Tim McNamara, FAHA, PhD, School of Languages and Linguistics, The University of Melbourne, Victoria, Australia
Gregory J. Meyer, PhD, Department of Psychology, University of Toledo, Toledo, OH
Joni L. Mihura, PhD, Department of Psychology, University of Toledo, Toledo, OH
Tyler M. Moore, Doctoral Candidate, Department of Psychology, University of California, Los Angeles
Bonnie Moradi, PhD, Department of Psychology, University of Florida, Gainesville
Kevin R. Murphy, PhD, Department of Psychology, The Pennsylvania State University, University Park
Jack A. Naglieri, PhD, ABAP, Curry Programs in Clinical and School Psychology, University of Virginia, Charlottesville; Professor Emeritus, Department of Psychology, George Mason University, Fairfax, VA; and Devereux Center for Resilient Children, Villanova, PA
Thomas Oakland, PhD, Department of Educational Psychology, University of Florida, Gainesville
Deniz S. Ones, PhD, Department of Psychology, University of Minnesota, Minneapolis
Mineko Anne Onoue, Doctoral Candidate, Department of Applied Psychology, New York University, New York, NY
Sheryl Packman, PhD, The College Board, New York, NY
Janet E. Panter, PhD, Department of Psychology, Rhodes College, Memphis, TN
Mike C. Parent, Doctoral Candidate, Department of Psychology, University of Florida, Gainesville
Yong Sue Park, PhD, Consortium for Multicultural Psychology Research, Michigan State University, East Lansing
David C. Parker, PhD, Department of Educational Psychology, University of Minnesota, Minneapolis

Randall D. Penfield, PhD, Department of Educational Research Methodology, University of North Carolina at Greensboro
Stephanie T. Pituc, Doctoral Candidate, Department of Psychology, University of Minnesota, Minneapolis
Robert E. Ployhart, PhD, Darla Moore School of Business, University of South Carolina, Columbia
John J. Prindle, PhD, Max Planck Institute for Human Development, Berlin, Germany
Antonio E. Puente, PhD, Department of Psychology, University of North Carolina, Wilmington
Antonio N. Puente, Doctoral Candidate, Department of Psychology, University of Georgia, Athens
Mark R. Raymond, PhD, National Board of Medical Examiners, Philadelphia, PA
Steven P. Reise, PhD, Department of Psychology and Advanced Quantitative Methods Training Program, University of California, Los Angeles
Kathleen T. Rhyner, Doctoral Candidate, Department of Psychology, University of Kansas, Lawrence
Carol Robinson-Zañartu, PhD, Department of Counseling and School Psychology, San Diego State University, San Diego
Michael C. Rodriguez, PhD, Quantitative Methods in Education, Educational Psychology, University of Minnesota, Minneapolis
Daniel E. Rohe, PhD, LP, Department of Psychiatry and Psychology and Department of Physical Medicine and Rehabilitation, College of Medicine, Mayo Clinic, Rochester, MN
Patrick J. Rottinghaus, PhD, Department of Psychology, Southern Illinois University, Carbondale
Ann Marie Ryan, PhD, Department of Psychology, Michigan State University, East Lansing
Jennifer L. Rymanowski, Doctoral Candidate, Department of Psychology, Syracuse University, Syracuse, NY
Paul R. Sackett, PhD, Department of Psychology, University of Minnesota, Minneapolis
Elizabeth D. Salmon, Doctoral Candidate, Department of Psychology, University of Maryland, College Park
Cynthia F. Salorio, PhD, Kennedy Krieger Institute; and Department of Physical Medicine and Rehabilitation, The Johns Hopkins University School of Medicine, Baltimore, MD
Juan I. Sanchez, PhD, Department of Management and International Business, Florida International University, Miami
Sylvia T. Scheuring, Doctoral Candidate, Arroki Inc., Lawrence, KS
Neal Schmitt, PhD, Professor Emeritus, Department of Psychology, Michigan State University, East Lansing
Katie L. Sharp, Doctoral Candidate, Department of Psychology, University of Kansas, Lawrence
Richard J. Shavelson, PhD, Professor Emeritus, School of Education, Stanford University, Stanford, CA
Stephen G. Sireci, PhD, Department of Educational Policy, Research and Administration and Center for Educational Assessment, School of Education, University of Massachusetts, Amherst
Finbarr C. Sloane, PhD, School of Education, University of Colorado at Boulder
Douglas K. Snyder, PhD, Department of Psychology, Texas A&M University, College Station, TX

Sueyoung L. Song, PhD, Department of Psychiatry, Dartmouth Medical School, Lebanon, NH
Michael F. Steger, PhD, Department of Psychology, Colorado State University, Fort Collins; and North-West University, Vanderbijkpark, South Africa
Steven E. Stemler, PhD, Department of Psychology, Wesleyan University, Middletown, CT
Robert J. Sternberg, PhD, Regents Professor of Psychology and Education, Oklahoma State University, Stillwater
Jennifer E. Stevenson, MPH, PhD, Department of Physical Medicine and Rehabilitation, The Johns Hopkins University School of Medicine, Baltimore, MD
Tia Sukin, EdD, Pacific Metrics Corporation, Monterey, CA
Lisa A. Suzuki, PhD, Department of Applied Psychology, New York University, New York, NY
Jane L. Swanson, PhD, Department of Psychology, Southern Illinois University, Carbondale
Elizabeth V. Swenson, PhD, JD, Department of Psychology, John Carroll University, University Heights, OH
Moin Syed, PhD, Department of Psychology, University of Minnesota, Minneapolis
Benjamin A. Tallman, PhD, Department of Physical Medicine and Rehabilitation, St. Luke's Hospital, Cedar Rapids, IA
Richard J. Tannenbaum, PhD, Educational Testing Service, Princeton, NJ
Jeanne A. Teresi, EdD, PhD, Columbia University Stroud Center; New York State Psychiatric Institute; and Research Division, Hebrew Home at Riverdale, New York, NY
George C. Thornton III, PhD, Professor Emeritus, Department of Psychology, Colorado State University, Fort Collins
Nancy T. Tippins, PhD, CEB Valtera, Greenville, SC
Adrea J. Truckenmiller, PhD, Department of Psychology, Syracuse University, Syracuse, NY
Tammi Vacha-Haase, PhD, Department of Psychology, Colorado State University, Fort Collins
Carly Vaughan, MSc, SHL Group Ltd., Surrey, United Kingdom
David A. Vermeersch, PhD, Department of Psychology, Loma Linda University, Loma Linda, CA
Scott I. Vrieze, PhD, Center for Statistical Genetics, University of Michigan, Ann Arbor
Michael E. Walker, PhD, Educational Testing Service, Princeton, NJ
Anna-Katherine Ward, Doctoral Candidate, Darla Moore School of Business, University of South Carolina, Columbia
Noreen M. Webb, PhD, Department of Education, Graduate School of Education and Information Studies, University of California, Los Angeles
Solange Muglia Wechsler, PhD, Department of Psychology, Pontifical Catholic University of Campinas, Campinas, Sao Paulo, Brazil
Irving B. Weiner, PhD, ABPP, Department of Psychiatry and Neurosciences, University of South Florida, Tampa
David J. Weiss, PhD, Department of Psychology, University of Minnesota, Minneapolis
Cathy Wendler, PhD, Educational Testing Service, Princeton, NJ
Edward W. Wiley, PhD, School of Education, University of Colorado, Boulder
Andrew Wiley, PhD, The College Board, New York, NY
Todd J. Wilkinson, PhD, Department of Psychology, University of Wisconsin–River Falls
Alexander J. Williams, Doctoral Candidate, Department of Psychology, University of Kansas, Lawrence
John O. Willis, EdD, SAIF, Rivier College, Nashua, NH

Hyung Chol Yoo, PhD, School of Social Transformation, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe

April L. Zenisky, EdD, Department of Educational Policy, Research and Administration and Center for Educational Assessment, School of Education, University of Massachusetts, Amherst

Michael J. Zieky, PhD, Educational Testing Service, Princeton, NJ

Bruno D. Zumbo, PhD, Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Vancouver, British Columbia, Canada

Series Preface

The *APA Handbook of Testing and Assessment in Psychology* is the ninth publication to be released in the American Psychological Association's *APA Handbooks in Psychology*™ series, instituted in 2010. The series primarily comprises multiple two- and three-volume sets focused on core subfields. Additionally, some single-volume handbooks on highly focused content areas within core subfields will be released in coming years.

The eight previously released sets are as follows:

- *APA Handbook of Industrial and Organizational Psychology*—three volumes; Sheldon Zedeck, Editor-in-Chief
- *APA Handbook of Ethics in Psychology*—two volumes; Samuel J. Knapp, Editor-in-Chief
- *APA Educational Psychology Handbook*—three volumes; Karen R. Harris, Steve Graham, and Tim Urdan, Editors-in-Chief
- *APA Handbook of Research Methods in Psychology*—three volumes; Harris Cooper, Editor-in-Chief
- *APA Addiction Syndrome Handbook*—two volumes; Howard J. Shaffer, Editor-in-Chief
- *APA Handbook of Counseling Psychology*—two volumes; Nadya A. Fouad, Editor-in-Chief
- *APA Handbook of Behavior Analysis*—two volumes; Gregory J. Madden, Editor-in-Chief
- *APA Handbook of Psychology, Religion, and Spirituality*—two volumes; Kenneth I. Pargament, Editor-in-Chief

Each set is primarily formulated to address the reference interests and needs of researchers, clinicians, and practitioners in psychology and allied behavioral fields. Each also targets graduate students in psychology who require well-organized, detailed supplementary texts, not only for "filling in" their own specialty areas but also for gaining sound familiarity with other established specialties and emerging trends across the breadth of psychology. Moreover, many of the sets will bear strong interest for professionals in pertinent complementary fields (i.e., depending on content area), be they corporate executives and human resources personnel; doctors, psychiatrists, and other health personnel; teachers and school administrators; cultural diversity and pastoral counselors; legal professionals; and so forth.

Under the direction of small and select editorial boards consisting of top scholars in the field, with chapters authored by both senior and rising researchers and practitioners, each reference set is committed to a steady focus on best science and best practice. Coverage converges on what is currently known in the particular subject area (including basic historical reviews) and the identification of the most pertinent sources of information in both core and

evolving literature. Volumes and chapters alike pinpoint practical issues; probe unresolved and controversial topics; and present future theoretical, research, and practice trends. The editors provide clear guidance to the "dialogue" among chapters, with internal cross-referencing that demonstrates a robust integration of topics to lead the user to a clearer understanding of the complex interrelationships within each field.

With the imprimatur of the largest scientific and professional organization representing psychology in the United States and the largest association of psychologists in the world, and with content edited and authored by some of its most respected members, the *APA Handbooks in Psychology* series will be the indispensable and authoritative reference resource to turn to for researchers, instructors, practitioners, and field leaders alike.

Gary R. VandenBos
APA Publisher

Introduction

As an undergraduate student, I was fortunate enough to hear a series of distinguished lectures at my college by a then recent past president of the American Psychological Association and a highly eminent psychological scientist, George A. Miller. I still remember a few of the points that he made in a couple of his lectures. In one, he stated that the two primary contributions of psychology to both science and the world were learning and measurement. Today, the concept of learning might be broadened to behavior change on the one hand and cognition and learning on the other, but his lecture was given in the days of Skinner and operant conditioning when the term *learning* was most appropriate. Nonetheless, that measurement was identified as an extraordinary and unique contribution of the field is significant. Psychologists focus on behavior; they sometimes attempt to change behavior, and to do so they need to know something about the base level of the behavior and subsequent resultant levels. These determinations involve measurement. That measurement has been seen as one of the most significant contributions of psychology lays the foundation for this handbook; indeed, measurement is one of the most critical cornerstones of the discipline. It cuts across all aspects of the reach of psychology.

E. L. Thorndike is often quoted as having stated, "Everything that exists, exists in some quantity and can therefore be measured."[1] That psychologists can measure many different behaviors in many different contexts is clear. If nothing else, these three volumes are proof of this point. That psychologists and other associated professionals have developed numerous quantitative approaches, often unbelievably complex and elaborate, to assigning numbers to the nature and amount of different behaviors, perceptions, thoughts, and feelings is also clear. It is unfortunate that those focusing on the psychological side of the measurement enterprise and those on the quantitative side often communicate so poorly with each other. My fantasy is that this handbook will increase communication among all of the participants in the testing and assessment process. At the professional level, this assembly includes psychometricians, both basic and applied psychologists, and researchers as well as test developers; test users; test administrators; and even, in some instances, test takers.

With such a formidable charge, perhaps the most daunting task during the initial stages of editing this handbook was organizing the volumes. I considered, for example, limiting the handbook to the work of psychometricians and keeping the volumes theoretical. I rejected

[1]What he actually said was, "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality" (Thorndike, 1918, p. 16). He was then misquoted by McCall (1939) as stating, "Whatever exists at all, exists in some amount." McCall went on to add, "Anything that exists in amount can be measured" (p. 15).

this approach because I happen to believe that the greatest value of testing is not the theoretical and scientific advances per se but the integration of testing and assessment into the field of psychology in all its varied manifestations.

A HISTORICAL INTRODUCTION

Not all who have considered measurement have seen it as entirely integrated into the field of psychology. Lee J. Cronbach's (1957) presidential address to the American Psychological Association literally characterized the science side of psychology as being divided into two subdisciplines, which he identified as experimental psychology and correlational psychology. Experimental psychology, of course, was regarded as so well known to the audience that Cronbach did not need to define or describe it. He believed, however, that he did need to describe correlational psychology, which he characterized as "slower to mature" (p. 671), more a study of relationships among variables occurring naturally in the world and an approach to research rather than statistical procedures per se. Correlational psychology was portrayed as naturalistic in view and in complete opposition to the high degree of control used by experimental psychologists. Correlational psychology had been identified, according to Cronbach, by a variety of other names, such as *ethnic psychology* (essentially cross-cultural psychology), *psychometric psychology, genetic psychology, comparative psychology,*[2] and *individual psychology*.[3] *Differential psychology* (e.g., Anastasi, 1958; Minton & Schneider, 1984) is a name that has been used to characterize the psychological side of measurement; it subsumes both the various constructs that psychologists measure and the names given historically to some of the preceding names, such as *individual psychology*. In this somewhat amorphous grouping of correlational psychology, Cronbach included developmental psychology, social psychology, personality psychology, and others. The theme of Cronbach's address was not to describe this historical fissure, however, but rather to call for the unification of scientific psychology as a type of confederation of branches of the field. Most applied psychologists today are correlational psychologists under Cronbach's classification scheme, whether such psychologists would use that term to describe themselves or not. The way they measure behaviors that are relatively naturally occurring and correlate the results of such measurements with other behaviors is what so identifies them. Correlational approaches differ from experimental approaches in that one does not control the behavior but rather attempts to keep constant or standardized the methods of gathering it, that is, collecting information in the same fashion from all test takers or participants.

Anastasi (1967), in her 1966 presidential address to the American Psychological Association's Division of Measurement and Evaluation, focused on a similar divide but one that was somewhat more focused on the testing community that was her audience. She feared that psychometricians, those psychologists focused on the mathematics supporting test construction and analysis, were becoming more distant from the psychological content that they were measuring. As she stated, "It is my contention that the isolation of psychometrics from other relevant areas of psychology is one of the conditions that have led to the prevalent public hostility toward testing" (p. 297). She continued,

> All . . . objections to psychological testing arise at least in part from popular misinformation about current testing practices, about the nature of available tests,

[2]Cronbach (1957) noted that the methods used in comparative psychology became more experimental and less naturalistic, and this discipline therefore left correlational psychology and merged with experimental psychology.

[3]Some of these distinctions have taken on other meanings in psychology.

> and about the meaning of test scores. Nevertheless, psychologists themselves are to some extent responsible for such misinformation. . . . Psychologists have contributed directly to the misinformation by actively perpetuating certain misconceptions about tests. (p. 300)

She went on to blame psychometricians for many of these concerns, stating, "Testing today is not adequately assimilating relevant developments from the science of behavior," and "Psychometricians appear to shed much of their psychological knowledge as they concentrate upon the minutiae of elegant statistical techniques" (p. 300). She observed that the dissociation between specialists in psychological testing and those in other areas of psychology keeps the content of many tests from incorporating the benefits of advances in psychological science. If these points were true in 1967 when the article was published, it is probably even more accurate today when statistical procedures and associated software have multiplied the complexities of analysis multifold. She concluded her article with the hope that psychological testing would "be brought into closer contact with other areas of psychology. Increasing specialization has led to a concentration upon the techniques of test construction without sufficient consideration of the implications of psychological research for the interpretation of test scores" (p. 305). To be sure, a goodly number of experts in testing have attempted to help psychometricians better understand the behavior that they are measuring (e.g., Carroll, 1976; Embretson, 1985; Lawrence & Shea, 2008).

A reader may wonder why I began the introduction to what I hope is an incredibly up-to-date handbook on testing and assessment in psychology with a brief summary of two articles that are about 50 years old, important though they may continue to be. The nature of the fractures described in these two articles is what, at least in part, formed the motivation for this handbook and justification for its organization; it is hoped that this handbook will show that at least some of these fractures have healed well.

ORGANIZATION OF THE HANDBOOK

Although psychological testing has proven invaluable in fostering psychological research on characteristics that are difficult to manipulate experimentally, where psychological testing has truly shown its value is in its application in a wide variety of manifestations. I spent considerable time over several weeks trying to choose among different ways to subdivide this handbook and decided to break the handbook into six primary areas. The first of these relates to the psychometric characteristics used to evaluate all measures, to analyze data emerging from measurements, and to research the constructs measured by these assessments. After this first section, the next five relate to traditional and historical areas of testing and assessment in psychology: industrial and organizational psychology, clinical psychology, counseling psychology, school psychology, and educational testing and assessment. This listing encompasses the most traditional applications of psychology. As in many fields, growth is often at the margins, and as readers will see, the associate editors and I have worked hard to address these emerging areas. Given that the chapters of this handbook are also available electronically on an individual basis, I could perhaps have had fewer sleepless nights over these decisions.

The first three sections, Test Theory, Types of Testing, and Industrial and Organizational Psychology, are found in the first volume. General Issues in Testing and Assessment in Professional Psychology, Clinical and Health Psychology, and Counseling Psychology are found

in the second volume. Finally, School Psychology, Educational Testing and Assessment, and Future Directions are found in the third volume. One could argue certainly for other arrangements, such as putting Test Theory, Types of Testing, and Educational Testing and Assessment together, but overall this clustering seems to work best. In fact, some studies have shown that the differences between clinical and counseling psychology are few or at least that they have considerable overlap.

As noted previously, the first section of the handbook covers psychometric aspects of testing. These aspects make up the quantitative underpinnings of testing and are applied, in various forms, in all applications of testing. Hubley and Zumbo provide a fine introduction to this section. After their chapter are chapters covering such traditional topics as reliability, validity, item analysis, equating and norming, test development strategies, factor analysis of test items and tests, the *Standards for Educational and Psychological Testing*, and evaluating tests themselves. Some newer topics are also included and represent fresher approaches to testing; these chapters include generalizability theory, item response theory, test fairness, the measurement of change, item banking, and ethical issues in testing. To be sure, many of those topics have been around for 50 years, but they rarely have much impact on introductory testing textbooks, for example. The chapters in the Types of Testing section relate to various types of testing; these chapters include material on objective testing and performance assessments in education, objective personality assessment, editing items to improve test quality, testing language skills, and fairness reviews of test items.

A section on testing and assessment in industrial and organizational psychology shares the first volume with the Test Theory and Types of Testing sections. This section has several introductory chapters, beginning with John P. Campbell's fine overview chapter, which outlines current and future practices. Other general chapters include those on work (job) analysis techniques, important individual cognitive difference variables, and performance appraisal. These chapters are followed by chapters relating to the types of variables used in making selection and placement decisions: biographical information, leadership, interviews, personality assessments, work samples, situational judgment measures, and holistic assessments. This section concludes with a couple of chapters on aspects of personnel selection in the modern world: multinational organizations and legal concerns in the job context. A few chapters also detail on-the-job work behaviors: counterproductive work behavior, stereotype threat, job satisfaction and other attitudes, and surveying workers.

The second volume provides chapters related to testing and assessment in clinical and counseling psychology. The first five chapters address general assessment in many areas of professional psychology, indeed, both clinical and counseling psychology. The introductory chapters provide an overview of assessment in clinical and counseling psychology, review the assessment process, describe methods of educating students in psychological testing principles and practice, consider clinical versus mechanical prediction, address communicating test results to clients, and discuss legal issues in clinical and counseling testing and assessment.

The chapters in the Clinical and Health Psychology section might have been organized into a few types of content. Several chapters detail the types of assessments that clinicians perform: the clinical interview, intellectual assessments, neuropsychological assessments, therapeutic outcome measures, and psychology and personality assessments, including performance assessment approaches (often known as projective measures). Several chapters describe contexts in which clinical testing takes place: treatment, adult mental health, child mental health, forensic, and medical settings, as well as multicultural testing.

The chapters describing testing and assessment in counseling psychology, as noted earlier, overlap in some contexts with those regarding clinical psychology, although we attempted to avoid outright duplication. These chapters could perhaps be rather cleanly broken into two groupings, those describing characteristics that are assessed and those describing counseling assessment contexts. The former includes chapters on the assessment of interests, career development; needs and values; self-efficacy; ethnic identity and acculturation; personality in counseling settings; racial stereotypes, discrimination, and racism; therapeutic assessment; gender-related attitudes and role identity; and meaning and quality of life. The latter describe the following contexts: rehabilitation counseling, occupational health settings, sport and exercise psychology, marriage and family counseling, custody hearings, and counseling with older adults. It should be clear that some of these are traditional to the earliest days of counseling psychology and others represent newer contexts.

The third volume includes testing and assessment in school psychology and education. Perhaps no career within professional psychology is more historically aligned with testing and assessment than school psychology. Of course, many school psychologists perform the same assessments as clinical and counseling psychologists, described in the chapters of the preceding volume. After the volume's introductory chapter, the vast majority of the chapters relate to specific types of assessments performed by school psychologists, including the assessment of preschool functioning; intellectual functioning; intellectual functioning using nonverbal measures; individual assessments of academic achievement; curricular assessment; dynamic assessment; behavioral, social, and emotional assessment of children; assessment of language competence; and assessment of adaptive behavior. A chapter comparing how assessments are performed in three countries is provided to demonstrate the cross-cultural nature of school psychology assessments. Perhaps unfortunately, no section on school psychological assessment would be complete without the requisite chapter on legal issues, and one such fine chapter is provided, too.

The Educational Testing and Measurement section is perhaps the most complex, and a few of the chapters in this section could perhaps have been included in the Test Theory and Types of Testing sections in Volume 1. Such overlap simply indicates how the various aspects of tests have consistencies regardless of context. This section begins with one of the most long-lived areas of educational testing, aptitude, higher education admissions, and outcomes assessment in higher education. Other contexts for educational testing include the K–12 context and licensure and certification testing. Many psychologists might be surprised to know that many more achievement tests are given each year in the schools than all the other psychological measures combined. A chapter is devoted to preparing students for taking tests. Two chapters are allotted to the testing of students with special needs and of English language learners; both are areas to which educational testing has devoted considerable attention in recent years. Several chapters relate to the use of tests in education. One of the hottest topics in educational testing as this handbook is being published relates to the evaluation of teachers using test data, and a chapter provides an up-to-date review of this literature. Other of these chapters relate to testing in educational and other contexts, especially industrial psychology. These contexts include setting standards on tests, developing and using multiple forms of tests, and reporting test score information in proper and communicative ways. Of course, this section would not be complete without a chapter on legal issues in educational testing.

A brief final section includes three chapters that look at future issues in testing. The first of these addresses the ever-growing practice of adapting tests from one language and culture to another. This initiative represents not only big business for the testing industry but also the expansion of psychology and psychological testing and assessment around the world. A fine chapter on test fairness is included in this section as well; it too is a chapter that could have been placed in the Test Theory and Types of Testing sections. Fairness is and needs to be so pervasive that I believe it needed to be included here as well as in those sections to demonstrate its criticality. John Hattie, currently president of the International Test Commission, and Heidi Leeson close the handbook with an ambitious chapter on the future of testing. I wanted that chapter to have an international focus, and I am glad they provided just such a perspective.

An internationally famous minister once told me that he was always depressed on Sunday afternoons. He would spend a week pouring himself into his sermon, only to see people mostly sleeping in church and not paying much mind to his words. One who engages in a task such as editing a three-volume handbook faces just such a fear. What I can report is that the authors of the chapters in this handbook represent an amazing array of talent. The handbook is a virtual hall of fame of living psychologists who have devoted themselves to the advancement of the field of psychology. I began this brief introduction with a loose quoting from former American Psychological Association President George A. Miller, who saw testing and assessment as one of the primary accomplishments and contributions of the field of psychology. Having read all 100 of the following chapters, I can certainly report that Miller was both correct and an accurate predictor of future events. Psychology is both a pure and an applied science, one that demonstrates its effectiveness and continually improves itself. Measurement and assessment play a huge part of that advancement and will do so in the future. Let me take this brief opportunity to thank the chapter authors not only for their chapters but for their persevering roles in advancing our field in a manner that both would make our psychological forefathers (and foremothers) proud and will lead to the betterment of humanity. No more, no less.

Finally, in addition to thanking the authors, I must express my very special thanks to the associate editors of this handbook. The six individuals who agreed to serve the field in this largely selfless manner also contributed to knowledge, to learning, and to the development of the field. They taught me; they corrected me; they identified chapter topics; they selected the outstanding authors whom I have just finished extolling. Professors all, Steven Reise led the efforts in the psychometric realm, Nathan Kuncel focused on the section relating to industrial and organizational psychology, Janet Carlson edited the clinical psychology section, Jo-Ida Hansen handled the counseling psychology section, Bruce Bracken covered the chapters relating to testing and assessment in school psychology, and Michael Rodriguez edited the chapters relating to educational testing. All six associate editors were truly partners. They all have my respect and gratitude for the rest of my life.

Finally, the publisher, the American Psychological Association, provided first-rate editors who helped me monitor the handbook's progress, provided motivation and support, and also served as partners in this enterprise. I hope that they will accept my thanks for their work well done.

Kurt F. Geisinger
Editor-in-Chief

References

Anastasi, A. (1958). *Differential psychology: Individual and group differences in behavior* (3rd ed.) New York, NY: Macmillan.

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist, 22*, 297–306.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27–56). Hillside, NJ: Erlbaum.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 11*, 671–684.

Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics*. New York, NY: Academic Press.

Lawrence, I. M., & Shea, E. (2008). *Improving assessment: The intersection of psychology and psychometrics* (ETS Research Memorandum 08-15). Princeton, NJ: Educational Testing Service.

McCall, W. A. (1939). *Measurement*. New York, NY: Macmillan.

Minton, H. L., & Schneider, F. W. (1984). *Differential psychology*. Long Grove, IL: Waveland Press.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In G. M. Whipple (Ed.), *The seventeenth yearbook of the National Society for the Study of Education: Part II. The measurement of educational products* (pp. 16–24). Bloomington, IL: Public School Publishing.

PART I

TEST THEORY

CHAPTER 1

PSYCHOMETRIC CHARACTERISTICS OF ASSESSMENT PROCEDURES: AN OVERVIEW

Anita M. Hubley and Bruno D. Zumbo

This chapter provides an overview of the psychometric characteristics of assessment procedures in psychology. *Psychometrics* is a field of study that focuses on the theory and techniques associated primarily with the measurement of constructs as well as the development, interpretation, and evaluation of tests and measures. A *construct* may be conceived of as a concept or a mental representation of shared attributes or characteristics, and it is assumed to exist because it gives rise to observable or measurable phenomena. *Measurement* is basically the description of attributes or characteristics in terms of numbers. In measurement, procedures and rules are applied to assign these numbers. It is important to remember, however, that the constructs that psychologists create, the tests or scales developed to measure those constructs, and the procedures and rules used in measurement are not value free; rather, each step in the process involves decisions that reflect personal, social, and cultural values as to what is important, useful, good, beneficial, and desirable—or not.

Assessment refers to the entire process of compiling information about a person and using it to make inferences about a person's characteristics or to predict behavior. An assessment involves combining and comparing information from a variety of sources such as interviews, records, observation, test results and information from other sources including family, friends, or professionals. Tests and measures are dominant assessment procedures in psychology. A test or measure may be thought of as a standardized procedure for sampling behavior. It can refer to a class test; a set of items or statements to which one responds, as with a questionnaire or interview; or a measure of reaction time, to give just a few examples. Whereas some might distinguish between tests and measures on the basis of whether there are correct responses (i.e., as in an educational achievement or knowledge test) or not (i.e., as in a personality or attitudinal measure), many use the terms interchangeably. Note also that whereas psychologists may commonly use the terms *test* and *measure,* particular fields or disciplines may prefer other terms such as *scale, instrument, questionnaire,* and *tool.*

As noted previously, psychometrics focuses on the theory and techniques associated with both the measurement of constructs and the development, interpretation, and evaluation of tests. In hopes of obtaining tests and measures that may be described as more reliable, valid, sensitive, and generalizable, the profession frequently introduces newly developed or revised psychological tests and a variety of techniques for evaluating them. Figure 1.1 provides a measurement and assessment framework to help readers visualize important psychometric elements reviewed in this chapter and addressed in more detail throughout the chapters in this section of the handbook.

ROLE OF THEORY

Figure 1.1 makes it clear that theory or theories play a role in the measurement and assessment framework. Multiple theories may come into play in this process. Some theories will be specific to the

DOI: 10.1037/14047-001
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

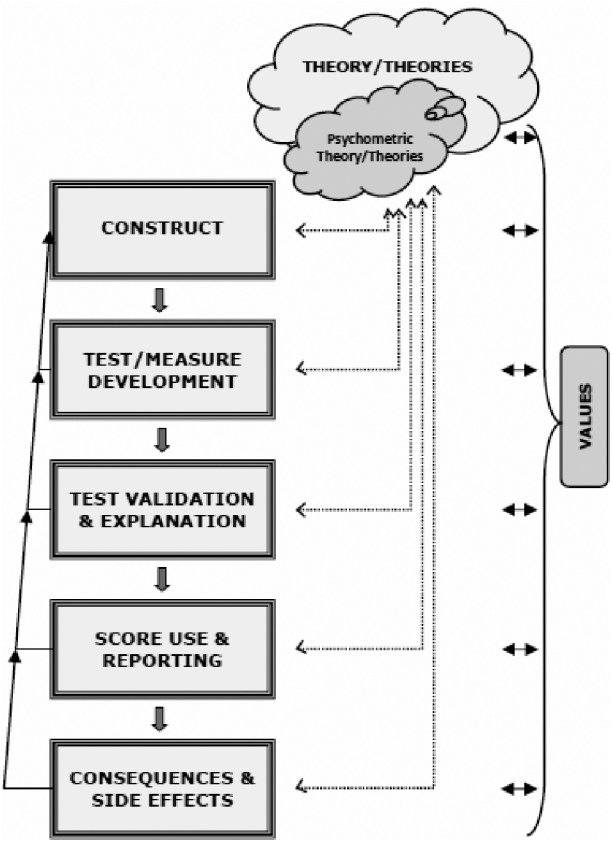

FIGURE 1.1. Measurement and assessment framework.

construct and content area of focus (e.g., self-concept theory), whereas others will be more general (e.g., social learning theory, life span developmental theory). Also of importance here is the role of psychometric theory (e.g., classical test theory, item response theory). The focus in this section is on introducing various psychometric theories, but it is important to remember that a variety of theories influencing different stakeholders in the process may play a role across the various elements of our framework.

There is not just one psychometric theory or one approach to measurement. Indeed, there are at least six (interrelated) measurement theories or classes of psychometric theory that are commonly referred to and may be grouped under observed score or latent variable approaches (Zumbo & Rupp, 2004). Observed score approaches include classical test theory (CTT; see Chapter 4, this volume) and generalizability theory (see Chapter 3, this volume). Latent variable approaches include factor-analytic theory (see Chapter 5, this volume), item response theory (IRT; see Chapter 6, this volume), Rasch theory, and mixture models. For many psychometricians, Rasch theory is a special case of IRT; however, as we describe later and in line with advocates of Rasch theory, we distinguish the two approaches.

CTT, or "true score" theory, is based on the decomposition of observed scores (X) into true (T) and error (E) scores: $X = T + E$. As the first, and still one of the most influential, of the measurement theories in psychology, CTT has been well described in classic textbooks by authors such as Lord and Novick (1968) and Allen and Yen (1979/2002). Note that the generic CTT statement, $X = T + E$, is axiomatic to other psychometric theories such as IRT and Rasch. Generalizability theory emerged in the 1970s and may best be viewed as an extension of CTT because it is often used to decompose the E in $X = T + E$ into different facets or sources (e.g., error resulting from items selected, raters used, gender of test administrator or examinee). Note that in unpacking the error, E, one implicitly redefines the true score, T.

Factor-analytic theory is the oldest of the latent variable theories, dating back to its first formal introduction in the early 1900s by Charles Spearman. Over the past century, factor-analytic theory has changed from being a descriptive psychometric approach characterized by a variety of principal components–based computational tools to an elaborated statistical modeling strategy that has a variety of model estimation methods and fit statistics. The move to considering factor-analytic theory within a statistical modeling framework and in a likelihood theory framework for model estimation and testing resulted in great advances from the 1960s to 1990s. The statistical formalization of factor-analytic theory has resulted in confirmatory modeling strategies and, most recently, a blend of confirmatory and exploratory approaches, referred to as *exploratory structural equation modeling* (Asparouhov & Muthén, 2009).

IRT emerged in the 1960s but, given the need for very large sample sizes in parametric IRT, only gained some popularity in psychology beginning in the 1990s. IRT focuses on the range of latent ability (theta) and the characteristics of items at various points along the continuum of this ability. That is, for example, whereas CTT produces only a single

estimate of reliability and standard error of measurement for a test, IRT has the advantage of producing these estimates across the range of the latent variable measured by a test.

Rasch theory can be mathematically characterized as IRT with only one item parameter, item difficulty. That is, one can think of Rasch theory as being a special case of three-parameter IRT, wherein the item discrimination and lower asymptote (sometimes called the *guessing parameter*) are fixed parameters and hence not estimated for the test at hand. One way of characterizing Rasch theory is that it has a guessing parameter of zero and an item discrimination parameter value of 1 for all items. Although mathematically correct, this description of Rasch theory does not capture the important philosophic orientation that comes with Rasch theory. In particular, Rasch theory often carries with it a belief that the model holds precedence over the data so that respondents or items are discarded until a Rasch model fits the remaining data.

Recent statistical developments have exploited the statistical interconnection among CTT, factor-analytic theory, and IRT under the rubric of latent variable modeling (Zumbo & Rupp, 2004). This more general form of modeling has resulted in two particularly useful advances. The first is that these latent variable models have incorporated the generalized linear modeling framework; hence, one can now more easily handle binary, ordinal, and other such data that are not continuous or multivariate normally distributed. Second, this general modeling has also incorporated strategies for mixture model theory, which allows for a more complex model to be fit to the data and allows for subsets of subjects in the population to have different parameterizations of that same general model. For example, under early forms of IRT, all of the respondents were treated interchangeably; therefore, all respondents had to be sampled from the same population. However, with mixture models, one can allow for specified subpopulations of respondents who might find an item easier (or more difficult) than others. This more flexible generalized mixture latent variable modeling allows for more complex models to be fit to data that reflect the more complex assessment scenarios researchers face on a day-to-day basis.

It is important to recognize that whenever researchers choose an approach to modeling data, they are implicitly imposing their values on that data (Zumbo, 2007; Zumbo & Rupp, 2004). For example, IRT modelers value characterizing subpopulation differences along the continuum of ability, whereas CTT modelers value fitting a model that is more universal. Rasch modelers, however, value the model over the data and so will remove data that do not fit the model. Indeed, proponents of Rasch theory believe that one does not even have meaningful measurement if the data do not conform to that one specific model.

BRIEF OVERVIEW OF TEST DEVELOPMENT STRATEGIES

Many different pieces are relevant to the development of scales and measures. In this section, only some of them are touched on, and test development approaches, types of tests, scaling, response formats, scoring, and item analysis are briefly reviewed.

Test Development Approaches

There are four primary approaches to the development of scales and measures: rational–theoretical, factor analytic, empirical criterion keyed, and projective (see, e.g., Martin, 1988). The rational–theoretical approach, in which the researcher uses either theory or an intuitive, commonsense approach to developing items for a test, is the most commonly used approach. In this case, expert opinion (i.e., of the researcher, a group of experts, a theory) forms the basis for the development and selection of items.

Probably the second most commonly used approach to test development is the factor-analytic approach. In this case, items are selected on the basis of whether they load on a factor, and a statistical rule forms the basis for the development and selection of items. Many large personality inventories (e.g., NEO Personality Inventory—Revised, Sixteen Personality Factor Questionnaire) have been developed using this approach. In addition, many tests today are developed using some combination of the rational–theoretical and factor-analytic approaches.

The empirical criterion-keyed approach, in which items are selected if they can discriminate the group of interest from a control group, is not

frequently used today in the development of measures. Several well-known measures were originally developed using empirical criterion-keyed approaches (e.g., Minnesota Multiphasic Personality Inventory, Strong Interest Blank).

Another approach to test development is the projective approach, although not many new tests are developed using this approach. The basic idea behind a projective test is to use ambiguous stimuli (e.g., inkblots, pictures) or have individuals create their own drawing (e.g., draw a person), and they will project their own concerns, fears, attitudes, and beliefs onto their interpretation or drawing. Projective tests are far less commonly used in North America than in other parts of the world (e.g., Europe).

Ozer and Reise (1994) described a powerful technique originally developed by Auke Tellegen and Niels Waller for test development and construct discovery. As Ozer and Reise wrote,

> Briefly, one begins with a rough idea of a personality construct and writes an overinclusive pool of possible items. Data are collected, and the analyses are used not just to refine the scale's psychometric properties, but also to generate new theories about the nature of the construct. A new set of possible trait indicators is then written, more data are collected and analyzed, and theory is again evaluated. This iterative process continues until a satisfactory level of convergence and demarcation of the construct has occurred and is manifested in the final set of items. (p. 368)

Types of Tests

Tests can be categorized in numerous ways. They may be categorized by field of study within psychology, for example, personality tests, intelligence tests, neuropsychological tests, interest inventories, achievement tests, aptitude tests, and behavioral tests. Tests may also be identified by their general administration procedures, that is, as individual tests that are administered one on one or as group tests that are administered to groups of individuals.

Another distinction has been made between tests of maximum performance and typical response. Tests of maximum performance measure how well an individual performs under standard conditions when exerting maximal effort and are presumed to include measures such as intelligence tests and achievement tests. Tests of typical response measure an individual's responses in a situation and are presumed to include measures such as personality tests and attitude scales. Tests may be further grouped according to the general type of information gathered. Specifically, tests may be (a) based on self-report (e.g., personality test, attitude measure, opinion poll), (b) based on performance or task (e.g., intelligence test, classroom test, eye exam, driver's test), or (c) observational (e.g., observation of play behaviors, observation in an interview). Combining these approaches, performance- or task-based tests may be seen as tests of maximum performance, whereas self-report and observational tests may be seen as typical response tests.

Another way in which tests have been categorized is as norm-referenced tests or criterion-referenced tests. These two types of tests differ in their purposes, manner in which content is selected, and the scoring process that defines how the test results must be interpreted. A norm-referenced test compares an individual's performance on a test with a predefined population or normative group, whereas a criterion-referenced test evaluates performance in terms of mastery of a set of well-defined objectives, skills, or competencies. In norm-referenced tests, items are generally selected to have average difficulty levels and high discrimination between low and high scorers on the test. In criterion-referenced tests, items are primarily selected on the basis of how well they match the learning outcomes that are deemed most important. Interpretation of test scores are based on percentiles, standard scores, and grade-equivalent scores in norm-referenced tests and on percentages or nonmastery–mastery categories in criterion-referenced tests. In norm-referenced tests, the normative value indicates how an individual scored relative to the normative group but provides relatively little information about the person's knowledge of, performance on, or level of the construct per se. Criterion-referenced test outcomes, however, give detailed information about how well a person has performed on each of the

objectives, skills, or competencies included in the test. A third type of test, ipsative, can be contrasted with norm-referenced tests. In ipsative tests, an individual's performance is compared with his or her performance either in the same domain or construct over time or relative to his or her performance on other domains or constructs. The latter case is sometimes referred to as *profiling*.

Scaling

Earlier, we stated that measurement is the description of attributes or characteristics using numbers and the procedures and rules used to assign these numbers. These procedures and rules are known as *scaling*. What one is trying to do with different scaling approaches is obtain a strong or faithful correspondence between the numbers and the attribute so that the numbers behave the way the attribute behaves. Thus, the challenges in scaling are the meaning of numbers, the ways in which the properties of numbers may be used to represent attributes, and the problems that arise in this process.

There are unidimensional and multidimensional scaling methods. The method used depends on whether one is attempting to compare people on one attribute or several. The most common unidimensional scaling methods are Thurstone's (1925) equal-appearing interval scaling, Likert's summative scaling (Likert, Roslow, & Murphy, 1993), and Guttman's (1944, 1950) scalogram analysis. Scaling methods may be used to scale stimuli, respondents, or both stimuli and respondents.

L. L. Thurstone (1925) developed the equal-appearing interval scaling method (commonly referred to as *Thurstone scaling*) in the context of attitude measurement. The idea behind this method was that one selects items that not only reflect a range of attitudes but also cover that range at roughly equal intervals. To do this, judges rate each item according to the severity of response (or level of attitude) it represented on an 11-point scale, and the mean and standard deviation (or median and interquartile range) are used to select items at these intervals. Respondents are asked to agree or disagree with each item; the score for each item is equal to the mean (or median) rating assigned to it, and the overall score is obtained by averaging the ratings over all the items with which the respondent agrees. Conducted properly, this expensive and time-consuming process is meant to produce scores on an interval scale. What is being scaled in Thurstone scaling are stimuli (i.e., items).

In 1932, Rensis Likert proposed a simpler technique called *summative scaling* (commonly referred to as *Likert scales*) that did not require judges to provide the ratings for items (see Likert et al., 1993). Respondents used symbols to indicate the degree to which they agreed or disagreed with statements, and these symbols were converted to a scale ranging, for example, from 1 to 5. The total score was obtained by summing the points assigned for each statement. What is being scaled in Likert scaling are respondents. The goal is to combine item responses for people in such a way that the obtained numbers (i.e., scores) represent reliable and valid individual differences among people. There has been disagreement over whether Likert's approach really does work as well as or better than Thurstone's more complex approach (see Drasgow, Chernyshenko, & Stark, 2010, and related commentaries).

Louis Guttman (1944, 1950) proposed an entirely different approach that he called *scalogram analysis* (commonly referred to as *cumulative scaling* or *Guttman scaling*). He pointed out that neither Thurstone's nor Likert's scaling methods were able to establish that the set of items on a test were unidimensional. He argued that being able to predict a respondent's entire set of responses to the items from the respondent's total score would demonstrate the presence of a unidimensional scale. A hypothetical, perfect Guttman scale consists of a unidimensional set of items that are ranked in order of difficulty from the least extreme position to the most extreme. Thus, theoretically, a person scoring a 3 on a five-item Guttman scale would agree with Items 1 to 3 and disagree with Items 4 and 5. This perfect relationship is rarely achieved, however, and thus some degree of deviation from this is expected (e.g., coefficient of reproducibility of .85 or more) before one decides that the model does not represent the attribute adequately. Many achievement tests may use a form of Guttman scaling to order questions on the basis of difficulty. Sometimes the examinee will be instructed to begin the test with a later item (e.g.,

Item 5). The assumption is that if the examinee can successfully answer items at that level, he or she would be able to answer the earlier items. Guttman scaling scales both respondents and stimuli and works best for constructs that are highly structured and hierarchical in nature.

Multidimensional scaling methods are appropriate to use when the attribute or phenomenon of interest involves many (sub)attributes or when one is investigating the structure of objects in multiple dimensions. *Multidimensional scaling methods* refers to a class of statistical techniques that explore similarities and dissimilarities (i.e., proximities) in data. These techniques produce a spatial representation of these proximities such that the more similar the data, the closer those points will be in the multidimensional space. The configuration is used to reveal the latent structure of the data. More important, in multidimensional scaling the distances between points in the multidimensional space should accurately reflect the proximities in the data. Several different types of multidimensional scaling exist (e.g., classical, metric, nonmetric; see Borg & Groenen, 2005).

Response Formats

The response formats used with psychological tests may be grouped into three categories: continuous, ordinal, or dichotomous. Common dichotomous response categories include yes–no, true–false, and agree–disagree. Continuous and ordinal response formats may be further divided into direct estimation and comparative methods.

Direct estimation methods. With direct estimation methods, the respondent provides a direct estimation of the magnitude of an attribute or characteristic. The most common ordinal response formats using direct estimation methods include semantic differential format, Likert-type format, and face–image format. A semantic differential format presents bipolar adjectives (e.g., *strong–weak*) as anchors with a number of points in between. In a Likert-type format, points along a continuum are labeled using either bipolar descriptors (e.g., *strongly disagree, disagree, neither disagree nor agree, agree, strongly agree*) or unipolar descriptors (e.g., *poor, fair, good, very good, excellent*). A considerable literature is available on the ideal number of points, the advantages and disadvantages of even versus odd numbers of points, what to label the midpoint on an odd-numbered response scale, and the meaning of and distance between various labels. Face–image format uses line drawings, usually of smiley faces (e.g., indicating a range of pain or satisfaction), without written descriptors or alongside either unipolar (e.g., *no pain* to *worst pain*) or bipolar (*very dissatisfied* to *very satisfied*) response descriptors. Face scales may be particularly useful when working with children, groups with cognitive challenges (e.g., patients with dementia, adults with intellectual impairments), or individuals with literacy difficulties.

The best known continuous response format using direct estimation is the visual analogue scale (originally called the *graphic rating method;* Hayes & Patterson, 1921), in which the respondent marks his or her response along a horizontal or a vertical 10-centimeter line with anchored endpoints. Typically, the score is the number of millimeters, ranging from 0 to 100.

Comparative methods. With direct comparative methods, the respondent is asked to compare the magnitude of an attribute or characteristic with something else. For example, individuals may be asked to rate their current quality of life relative to (a) other people, (b) their quality of life in the past, or (c) their ideal quality of life. A Likert-type response scale is most commonly used with comparative methods (e.g., ranging from *much worse than others* to *neither worse nor better than others* to *much better than others*). With indirect comparative methods, respondents are typically asked to rate the magnitude of an attribute or characteristic for themselves and for someone else. For example, individuals may be asked to rate, separately, their current quality of life and the average person's quality of life. The two different scores are then compared.

Scoring

Most item responses to scales and measures may be scored to produce continuous, ordinal, or dichotomous scores. Continuous scores consist of an infinite or very large number of points. Items measuring height and temperature usually produce scores that are treated as continuous. Some researchers will

treat a fairly large number of response points (e.g., 10-point Likert-type format) as practically continuous scores, whereas others will treat them as ordinal scores. Ordinal scores (also known as *graded response* or *ordered polytomous scores*) typically involve three to 10 possible values (e.g., from a 5-point response format). Dichotomous or binary scores consist of only two values (e.g., 0 or 1) and may be obtained from responses that are (a) rated and scored dichotomously using a scoring key (e.g., true–false, agree–disagree, yes–no) or (b) rated using more than two response options but scored dichotomously as correct or incorrect (e.g., as in a multiple-choice item) or present or not present (e.g., as in a screening or diagnostic test item).

Reverse scoring. Use of a mix of positively and negatively worded items on a scale or measure has historically been encouraged to identify when respondents are displaying acquiescence (i.e., the tendency for respondents to generally agree with items) or are not paying attention to the content of the individual items. Before aggregating items to obtain a composite score, one set of items (either the positive or the negative ones, as appropriate to the particular measure) must first be reverse scored. When reverse scoring an item, the values on the scale are reversed (e.g., on a 5-point scale, 5 = 1, 4 = 2, 3 = 3, 2 = 4, and 1 = 5 for the reverse-scored items). Despite the original reasons for including a mix of positively and negatively worded items on a scale, more recent research has demonstrated problems with some respondents' understanding of negatively worded items, response distributions, factor structure, and reliability estimates (see, e.g., Barnette, 2000).

Scale or total scores. When psychological scales or measures consist of multiple items, these items are scored and aggregated in some way to obtain a composite scale or total score. This composite score may be the sum or the average of the item scores. Typically, this composite score will be treated as a continuous score (e.g., ranging from 0 to 63), but depending on the range of the composite score, it could be ordinal (e.g., ranging from 1 to 7) or dichotomous. Note that the properties of the scores that will be used in any analysis (not the response format) is what is important here.

Item weighting. Most psychological scales use summed composite scores in which equal weight or value is assigned to each item. It has been argued, however, that some items may be more important to the construct or to the respondent than other items and thus should be assigned more weight in calculating the composite or total score.

There are two types of weighting. *External* or *objective weights* are determined in advance on a theoretical or statistical basis and applied equally to all respondents before calculating composite scores. *Internal* or *subjective weights* reflect each individual respondent's evaluation of which domains or items have greater importance; in this case, each respondent's composite score is based on his or her own unique set of weights. Although different weighting schemes have been proposed, the most commonly used approach is to use multiplicative scores in which item importance ratings and item scores are multiplied to obtain weighted scores, which are then summed. Despite the intuitive appeal of weighted scores, the empirical evidence does not suggest that they provide any improvement over unweighted scores (see, e.g., Russell & Hubley, 2005).

Item Analysis

Item analysis can be used both in the test development process to aid in item revision and later to help understand why a test shows specific levels of reliability and validity. Different item analysis indicators are used in CTT and IRT.

Classical test theory. Some common CTT item analysis indicators include alpha-if-item deleted, item–total correlations and corrected item–total correlations, interitem correlations, item difficulty index, and item discrimination index. Alpha-if-item-deleted values indicate the Cronbach's alpha for the scale if an item is discarded. This information is provided for each item of a scale. A notably higher alpha-if-item-deleted than the overall Cronbach's alpha suggests that the item should be dropped or revised because the scale appears to function better without that item. Item–total correlations consist of the correlation between the score on a single item and the composite or total score on the scale, whereas corrected item–total correlations consist of

the correlation between the score on a single item and a composite score that does not include that single item. These values are used to identify problem items that show negative or near-zero correlations with the composite score. The interitem correlation matrix is used to help understand low (corrected) item–total correlations. Patterns, such as an item (a) not being correlated with many (or even any) of the other items on the test and (b) showing positive correlations with some items but negative or zero correlations with other items, suggest one is tapping either some other aspect of the construct that is not well represented by the items in the test or another construct altogether. The item difficulty index is the proportion of respondents who answer an item correctly. The item discrimination index is a measure of the effectiveness of an item in discriminating between high and low scorers on a test. Generally, the item discrimination index is maximized when the item difficulty index is close to .5.

Item response theory. In IRT, common item analysis indicators include difficulty (the *b* parameter), discrimination (the *a* parameter), guessing (the *c* parameter), conditional reliability, and conditional standard error of measurement. Unlike CTT, which considers item difficulty and discrimination in relation to the sample of respondents, IRT examines difficulty and discrimination across the range of the latent variable. The *item characteristic curve* is the regression of the probability of endorsing the item (or, in achievement tests, the probability of getting the item correct) onto the latent variable score, which is commonly denoted as theta.

Describing the item parameters in more detail, the higher the *b* parameter is, the more difficult the item. One can look at the difficulty of an item across the latent ability range, or one can compare difficulty across items at different points in the latent ability range. As previously noted, in a one-parameter IRT model, only the *b* parameter varies (one variant of the one-parameter model is called the *Rasch model*). Discrimination is identified by the slope of the item characteristic curve at its steepest point. The steeper the curve and the larger parameter *a* is, the more discriminating the item. As the *a* parameter decreases, the curve gets flatter until there is virtually no change in probability across the latent ability continuum. Items with very low *a* values are not able to distinguish well among people with varying levels of latent ability. The two-parameter IRT model allows both *a* and *b* parameters to vary in describing the items. Guessing is indicated by the *c* parameter or lower asymptote—it is the probability of selecting the correct answer at the lowest level of ability. The higher the *c* parameter is, the greater the probability of guessing or selecting the correct answer, even though the latent ability is low. The *c* parameter is often used to model guessing on multiple-choice items. The three-parameter IRT model allows the *a*, *b*, and *c* parameters to vary to describe the items. Finally, whereas CTT provides a single reliability estimate and standard error of measurement for a scale or measure, IRT uses a graph showing the item information function to display these values (called *conditional reliability* and *conditional standard error of measurement*) for all points across the latent ability range (see Chapter 6, this volume).

Role of Content Validation, Factor Analysis, and Reliability

Content validation, factor analysis, and reliability estimates are often used in the test development process. However, because each of these also provides critical validity evidence, they are described in detail in the next section. Sireci and Sukin (Chapter 4, this volume) also provide an overview of some commonly used validation methods. It is worth noting that formulas exist for (a) determining how many more items are needed or how many times longer a test with a given reliability must be to attain a desired (usually higher) reliability and (b) estimating what the reliability might be if one added more items; this information can be of great assistance in revising measures.

TEST VALIDATION AND EXPLANATION

Validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). It involves presenting evidence and providing a

compelling argument to support the intended inference and show that alternative or competing inferences are not more viable. Zumbo (2009) took this definition a step further to argue that traditional validation practices (e.g., factor structure, reliability coefficients, validity coefficients) are descriptive rather than explanatory and that validity should, in addition, provide a richer explanation for observed test score variation. There is an important difference between validity and validation. *Validation* is about the process or methods used to support validity and an explanation for score variation. It should also make explicit any personal or social values that overtly or inadvertently influence that process. Thus, validation is an ongoing process that, as Messick (1989) noted, "is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what test scores mean" (p. 13).

Unified Model of Validity

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) has endorsed a unified model of validity that replaced older views in which validity was seen as a property of a test, was supported by different types of validity (i.e., content, criterion, construct), and involved a dichotomous decision (i.e., valid or invalid; Hubley & Zumbo, 1996). Messick (1989) presented a progressive matrix of validity to emphasize the evidence that is needed when validating inferences from tests. The matrix is organized in terms of the basis for justifying validity (i.e., evidential basis vs. consequential basis) and the test function (i.e., test interpretation vs. use). Under the evidential basis, (construct) validity evidence is needed to support a given test interpretation and evidence of the relevance and utility of the test score inferences and decisions is also required to support the use of test scores. Most unique to Messick's (1989) progressive matrix is the consequential basis, which adds both value implications and social consequences. Value implications challenge researchers to reflect on the values that led to their interest in and labeling of the construct, the theory underlying the construct and its measurement, and the broader social ideologies that affected the development of the identified theory (Messick, 1980, 1989). *Social consequences* refer to the unanticipated or unintended consequences of legitimate test interpretation and use (Messick, 1998). Despite the misconceptions rampant in the literature, the concept of social consequences, as described by Messick (1998), does not include test misuse. Messick (1998) was most concerned about the relationship between score meaning and social consequences. If social consequences occur that are traceable to construct underrepresentation or construct-irrelevant variance, then they are considered a form of validity evidence; if they are not, then they are not part of validity (Messick, 1998). Whereas most test developers and users seem to agree that social consequences are important considerations in testing, not everyone supports the inclusion of social consequences in validity and validation (e.g., Brennan, 2006; Popham, 1997).

Hubley and Zumbo (2011) presented a framework for test validation that expands on the concept of consequences while also putting them in their proper place relative to other types of validity evidence. This framework is highlighted in Figure 1.2. Specifically, forms of evidence that may be presented to support the interpretation and use of test scores include, but are not limited to, score structure; reliability; content-related evidence; criterion-related evidence; convergent and discriminant evidence; known-groups evidence; generalizability or invariance evidence across samples, contexts, and purposes; intended social and personal consequences; and unintended social and personal side effects. In this framework, it is also important to note that (a) the arrows from the various forms of evidence, the intended social and personal consequences, unintended social and personal side effects, and test score meaning and inference are bidirectional, implying, for example, that unintended social and personal side effects can have an impact on test score meaning and inference, and vice versa, and (b) psychometric and other theories as well as values influence the construct, the measure, and validity and validation. Some of this evidence is briefly described next.

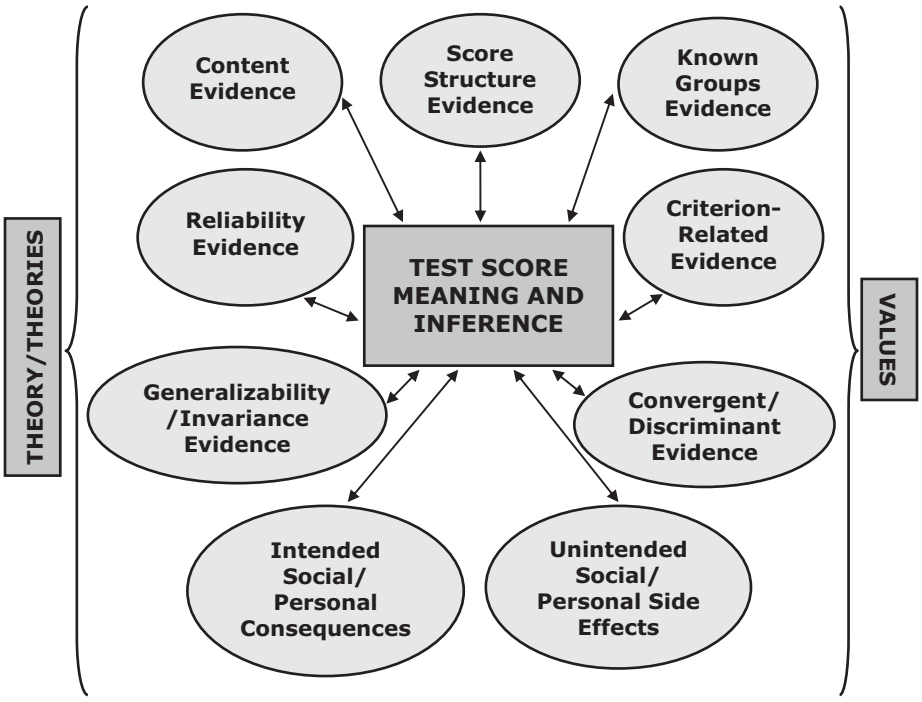

FIGURE 1.2. Test validation framework.

Content-related evidence. Content-related evidence of validity examines the "degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose" (Haynes, Richard, & Kubany, 1995, p. 238). In a content validation study, subject matter experts evaluate elements of a test or measure and rate them according to their relevance and representativeness to the content domain being tested. Haynes et al. (1995) made the important point that all elements of a test, including not just item content but also instructions, response formats, scoring instructions, and so forth, be subjected to content validation. It is important that a test developer creates a conceptual and operational definition of the construct and subject it to content validation before developing other elements of the instrument. Depending on what one wants to evaluate, using different experts may be advantageous. For example, experts in the field might examine all elements of the measure; those who will be administering the instrument might examine item content, administration instructions, and scoring instructions; and the target population (whom we refer to as *experiential experts*) might examine the item content, response format, and layout. By using subject matter and experiential experts for initial generation of items and other elements, one can help ensure that items and other elements are relevant to, and representative of, the construct. Obtaining both quantitative evidence of content validity and qualitative feedback is useful. In terms of quantitative evidence, a variety of indices are used to judge interrater agreement in content validation, but the most common one is the content validity ratio (Lawshe, 1975) or content validity index (Lynn, 1986). Finally, one should also examine the proportional representation of items in the measure. Items should be included in a way that reflects the relative importance of each facet of the construct.

Score structure. Factor analysis (see Chapter 5, this volume) is a statistical method used to (a) discover how many factors (representing the latent variables) are being tapped by the items in a test (i.e., exploratory factor analysis) or (b) confirm whether the test items measure the factors as intended (i.e., confirmatory factor analysis). Knowing the factor structure of tests is important because it greatly affects how one scores a test and how one assesses the reliability of scores and the validity of inferences made from a test. Specifically,

if a test is shown to have a unidimensional (i.e., single-factor) or essentially unidimensional (i.e., predominantly single-factor) structure, then the responses to items can be summed to form a composite score, and that score can be used in any psychometric evaluation of reliability and validity. If, however, a test is shown to have a multidimensional (i.e., two or more factors) structure, then responses to items that load on different factors must be summed to form subscale scores. When computing reliability coefficients, separate coefficients must be obtained for each subscale. Likewise, when evaluating the validity of inferences, analyses are conducted separately for each subscale. Unfortunately, it is not uncommon for factor analyses to show a multidimensional structure for a test, and yet researchers and practitioners continue to use total scores. One should not use total scores or report overall internal consistency or test–retest reliability coefficients unless factor structure evidence has been provided to support a total score, such as when all the factors are highly intercorrelated. In the case of a multidimensional test, one would need to conduct a higher order factor analysis to support the use of a total score in addition to subscale scores; it is not enough to show that the factors are highly correlated.

Reliability. *Reliability* refers to the degree to which test scores are repeatable or consistent. Alternatively, one can think of reliability as the extent to which test scores are free from measurement error. There are three basic ways of conceptualizing the reliability of scores that produce a variety of reliability estimates or coefficients. An additional set of reliability estimates focuses on the consistency of scores or a scoring system when different raters are used.

Viewing reliability as equivalence, one can correlate scores from two different forms of the same test that were administered to the same group of respondents in the same session or on different occasions, which is referred to as *alternate forms reliability*. Viewing reliability as stability, one can correlate scores from the same test administered on two different occasions (i.e., test and retest) to the same group of individuals, which is referred to *as test–retest reliability*. It is important to select a time interval that is not so short that respondents will recall their responses to items but even more important to select one that is not so long that one would expect changes in the construct to occur for respondents.

When viewing reliability as internal consistency, three different estimates of reliability (split-half reliability, coefficient alpha, Kuder–Richardson formula 20) can be used. In split-half reliability, one correlates the scores from two halves of a test that has been administered only once and applies the Spearman–Brown formula to correct for an underestimation of reliability that results from treating the test as being only half its length. Cronbach's coefficient alpha is the most commonly reported estimate of reliability. It can be thought of as the mean of all possible split-half coefficients corrected by the Spearman–Brown formula. More recently, ordinal alpha has been introduced, which provides a more accurate internal consistency estimate when used with items with ordinal response formats (Zumbo, Gadermann, & Zeisser, 2007). The Kuder–Richardson formula 20 refers to the case in which coefficient alpha is used with dichotomous (i.e., yes–no, true–false) data. In each case, a reliability coefficient tells you the variability in the sample scores that is the result of individual differences rather than random (unsystematic) measurement error.

In interrater (or interscorer) reliability, one is interested in how repeatable the scores are when two or more different people are scoring or observing the same behavior. If interrater reliability is low, then one needs to consider whether the scoring criteria are clear or complete enough, whether training of the raters was insufficient, or whether some raters are not doing a good job. Commonly used interrater reliability estimates include consensus estimates such as percentage of agreement and Cohen's kappa and consistency estimates such as Pearson's product–moment correlation coefficient, intraclass correlations, and Cronbach's alpha.

Reliability and validity coefficients provide related but separate information. High reliability coefficients indicate a high proportion of true score variance in the observed scores, and thus one can feel confident that one is measuring real individual differences. Reliability is commonly viewed as a necessary but insufficient condition for validity. Thus, although it is important that a measure be reliable

and produce repeatable or consistent scores, it does not mean that the inferences one wants to make are valid.

Criterion-related evidence. Criterion-related evidence demonstrates the degree to which scores obtained on a measure are related to a criterion. A criterion is an outcome indicator that represents the construct, diagnosis, or behavior that one is attempting to predict using a measure. One can think of the criterion as being what one would really like to have (e.g., diagnosis) but what one often cannot obtain because of cost or time, and so the measure (e.g., screening test) acts as a substitute or shortcut. The value of a criterion-related study is dependent on both the quality of the criterion selected and the validity of the inference made from that criterion. A criterion validity coefficient consists of the correlation between the score on a measure and the criterion. The larger the coefficient is, the better the evidence provided. Criterion-related evidence can be described as either predictive or concurrent. *Predictive evidence* examines how well a score on a measure is related to or predicts a future criterion (i.e., a behavior, test performance, or diagnosis obtained at a later date), whereas *concurrent evidence* examines how well a score on a measure is related to or predicts a current criterion (i.e., a behavior, test performance, or diagnosis obtained at the same time or nearly the same time). Another way to look at the relationship between the measure and criterion is to examine the standard error of estimate, which provides the margin of error to be expected in the predicted criterion score. The standard error of estimate ranges from 0 to the value of the standard deviation of the criterion score.

Criterion-related validity evidence may be further evaluated using Taylor–Russell tables in some areas of psychology (e.g., industrial and organizational psychology) when making decisions using test information. Taylor and Russell (1939) argued that there are three important factors when judging predictive validity: criterion-related validity, base rate, and selection ratio. The *base rate,* in essence, reflects how often something normally occurs, and the *selection ratio* is the proportion of individuals in the relevant population (e.g., applicants) who are selected. All other things being equal, (a) the larger the criterion-related validity coefficient is, the more useful the test will be; (b) tests are most useful when the base rate of success is .50; and (c) the smaller the selection ratio is, the more useful the test will be.

In areas such as clinical psychology, sensitivity, specificity, and predictive values are all used as evidence of criterion-related validity. As an example of these terms, consider using depression as the construct of interest. Sensitivity = true positives / (false negatives + true positives) and identifies the percentage of depressed people in the sample (based on the criterion) that the depression scale correctly identified as depressed. Specificity = true negatives / (true negatives + false positives) and indicates the percentage of nondepressed people in the sample (according to the criterion) that the depression scale correctly identified as nondepressed. The positive predictive value = true positives / (false positives + true positives). It shows the percentage of individuals who are truly depressed (according to the criterion) out of those whom the scale identified as depressed. Finally, the negative predictive value = true negatives / (true negatives + false negatives) and indicates the percentage of people who are truly not depressed (according to the criterion) out of those whom the scale identified as nondepressed. Receiver operating characteristic curves graph sensitivity and specificity for all of the possible scores of a scale, which is achieved by plotting true positives (sensitivity) on the vertical axis and false positives (1 − specificity) on the horizontal axis. The more clearly a scale is able to discriminate between, for example, depressed and nondepressed individuals, the greater the curve will deviate from the (straight) line of no information toward the upper left corner of the graph. The *area under the curve* is an estimate of the probability that a randomly chosen depressed person will have a higher test score than a randomly chosen nondepressed person. The line of no information has an area-under-the-curve probability of .50, whereas a perfect test would have an area-under-the-curve probability of 1.00. Calculating the standard error of the area under the curve tells one whether the area under the curve for a test is significantly different from the line of no information, that is, whether the test provides one with any more information than not administering the test.

Convergent and discriminant evidence. Convergent measures may consist of measures of highly related constructs (e.g., depression and anxiety) or the same constructs (e.g., depression); in the latter case, correlations of such scores are sometimes misidentified as criterion-related validity evidence. Discriminant measures may consist of theoretically unrelated constructs (e.g., depression and intelligence) or constructs between which one wants to distinguish (e.g., depression from anxiety). Convergent and discriminant measures may be considered to be mapped on a continuum; it can sometimes be difficult to pinpoint when a measure is a convergent measure or a discriminant measure because it is a matter of degree rather than category. Sometimes knowing whether a measure is convergent or discriminant is important (e.g., when determining whether a depression measure is more related to measures of depression than anxiety), and sometimes it is not (e.g., when the pattern of relationships is more important than the label of convergent or discriminant). Correlations between convergent measures should be relatively high, whereas correlations between discriminant measures should be relatively low. Most important, discriminant validity coefficients should be significantly lower than convergent validity coefficients.

The terms *convergent validity* and *discriminant validity,* and a more conceptually sophisticated methodology called the *multitrait–multimethod matrix* to assess them both in a single study, were introduced by Campbell and Fiske (1959). This methodology requires that one measure each of several constructs (or traits) by each of several methods (e.g., paper-and-pencil test, observation). The multitrait–multimethod matrix provides a matrix for organizing obtained reliability and validity coefficients. Generally, one expects that reliability coefficients will be higher than validity coefficients. Convergent validity evidence is provided if monotrait–heteromethod correlations are notably higher than any other validity coefficients. Discriminant evidence is provided by both heterotrait–monomethod and heterotrait–heteromethod correlations, but the strongest evidence comes from the latter group of correlations because they share neither construct nor method. A key strength of the multitrait–multimethod matrix is that the impact of the method used to measure a construct on the magnitude of convergent and discriminant validity coefficients may be explicitly evaluated. A methods effect is seen to be present if heterotrait–monomethod correlations are notably higher than heterotrait–heteromethod correlations.

Validity and Multilevel Measures

In psychology, multilevel constructs and tests are increasingly being used in assessment and evaluation. A multilevel construct refers to "a phenomenon that is potentially differentially meaningful both in use and interpretation at the level of individuals and at one or more levels of aggregation" (Zumbo & Forer, 2011, p. 177). The level at which the test scores are interpreted has important consequences for validation. When scores from multilevel tests and measures are interpreted, used, and reported at an aggregate or group level (e.g., at the level of school, neighborhood, or country), validity evidence also needs to reflect this same group level of data (Forer & Zumbo, 2011; Zumbo & Forer, 2011). Thus, one must report, for example, reliability, factor structure, and validity evidence gathered at the appropriate level. Moreover, one must consider the potential errors in inference that may occur across levels of data (i.e., ecological or atomistic fallacies of measurement data inferences; Zumbo & Forer, 2011).

Score Use and Reporting

A test or measure is a standardized procedure for sampling behavior. Standardization means that the test is administered to each person using the same materials, instructions, and scoring procedures. Standardization is important to ensure, as much as possible, that differences seen in performance on a test are due to individual differences on the construct of interest and not due to random error associated with how the test was administered or scored. A *standardization sample* is a large group of individuals who are representative of the target population for whom the test is intended. The standardization sample provides some indication of the scores that a particular group might obtain if a given standardization procedure is followed. There can be more than one standardization sample. When the test is intended

for use with "everyone" (e.g., U.S. adults), a test developer may strive to have a standardization sample that represents the national population and matches important census data (e.g., gender, race or ethnicity, geographic distribution).

In psychological research, one might report descriptive information (e.g., means, standard deviations) and use raw scores in the calculation of inferential statistics. In clinical work, however, raw scores by themselves are often meaningless, and making use of normative scores or norms becomes important. Norms serve as a frame of reference for interpreting raw scores and allow researchers to compare people by indicating a person's standing on a test relative to the distribution of scores obtained by people of the same chronological age, grade, sex, or other demographic characteristics. By determining the distribution of raw scores by the standardization sample on the test, researchers can then convert the raw scores to some form of standard scores or norms. If the scores of a standardization sample are converted to standard scores and thus provide norms, then this group is also known as a *normative group*.

It is important to remember that populations change over time, and thus a normative sample may not reflect the population very well over time or the performance obtained by a normative sample may no longer accurately reflect the range of performance by the population. As a result, it is important that norms be updated over time. Recognizing the importance of standardization when using norms is also key. There may be occasions when one chooses not to standardize administration and scoring of a test (e.g., when one wants to test the limits and determine whether an individual could complete a task correctly if the standard time limits were loosened). However, one must then recognize that this person's score should not be compared with any norms that are based on standardized procedures that one did not use.

There are a variety of different types of norms, but the two types of norms used most commonly are percentiles and standard scores. The most common type of raw score transformation in psychological and educational testing is percentiles. A *percentile* is a ranking that provides information about the relative position of a score within a distribution of scores. More specifically, a percentile indicates the percentage of people in the specific normative group whose scores fall below a given raw score. Cumulative percentiles indicate the percentage of people in the specific normative group whose scores fall at or below a given raw score. Cumulative percentiles are particularly useful with data that are highly skewed, such as depression or self-esteem scores. A major drawback of percentiles is that they maximize differences in the center of the distribution and minimize differences between raw scores in the tails of the distribution because the highest frequency of scores occurs in the middle of the distribution. Thus, small differences in the raw scores at the center of the distribution result in large differences in the percentiles. The opposite effect occurs in the tails of the distribution.

Standard scores indicate where a raw score sits in the distribution relative to the mean. Specifically, they indicate (a) how far away the raw score is from the mean and (b) whether the raw score is above or below the mean. Common standard scores are *z* scores, with a mean of 0 and a standard deviation of 1, and *T* scores, with a mean of 50 and a standard deviation of 10. Standard scores are simple linear transformations of raw scores. That is, the mean and standard deviation of the scores may change, but the shape of the distribution remains exactly the same. This also means that one does not get the distortion that is present with percentiles. Standard scores may also allow one to compare an individual's performance on two or more different tests. However, if the distribution of scores on a measure is not normal, then one cannot make certain statements that a normal distribution allows (e.g., 95% of the scores are within 2 standard deviations above and below the mean), and one cannot compare scores on two tests properly if the shape of the two distributions is different. Distributions can be normalized by conducting a nonlinear transformation (to essentially stretch the skewed curve into the shape of a normal curve) and then calculate standard scores. Such scores are referred to as *normalized standard scores*. One common example of a standard score based on a nonlinear transformation is the *stanine,* which has a mean of 5 and a standard deviation of approximately 2. It is important to note that means and standard deviations

presented by group (e.g., by gender, age, or grade) are not norms. They are simply descriptive statistics.

Norms consist of percentiles and standard scores. Sometimes a normative group is treated as one single group (e.g., adults), but norms are more typically presented by specific subgroups. That is, norms are often used to interpret an individual's performance relative to that of other people of the same age group, education or grade level, gender, or race or ethnicity. These four variables are most commonly used to form norms, but norms can be based on any number of demographic variables.

Item and Construct Mapping

Thurstone (1925), in the context of measuring intellectual growth and scaling across ages or grades, and Ebel (1962), in the context of educational testing, were among the first to suggest that test scores should be interpreted in terms of characteristic or representative items rather than norms. Ebel described the fundamental principle quite nicely when he wrote,

> Unfortunately, something important tends to get lost when raw scores are transformed into normative standard scores. What gets lost is a meaningful relation between the score on the test and the character of the performance it is supposed to measure. It is not very useful to know that Johnny is superior to 84 per cent of his peers unless we know what it is that he can do better than they, and just how well he can do it! (p. 18)

Both Thurstone (1925) and Ebel (1962) proposed that score meaning and interpretation be based on a map of the scores (or ranges of scores) for the items in a test. Thurstone's and Ebel's ideas have been expanded and further elaborated using item mapping (e.g., Bock, Mislevy, & Woodson, 1982), which makes use of exemplar items to characterize particular score points on educational assessments. This concept has been further elaborated for both educational and psychological testing into a construct map (Wilson, 2003, 2004). Both item mapping and construct mapping have been developed in the context of Rasch and IRT models. These models provide users with latent variable (or factor) scores, often denoted theta, that are on a scale with a specified mean and variance (e.g., *z* or *T* scores). Item or construct mapping has been a very useful strategy for helping test developers provide meaning for Rasch or IRT scores by mapping scores (or bands of scores) to, for example, typical item responses.

Consequences and Side Effects of Using Tests and Measures

Intended consequences and unintended side effects of legitimate test use appear both as a separate stage in the measurement and assessment framework (see Figure 1.1) and as validity evidence (see Figure 1.2). This placement highlights that these consequences and side effects occur naturally as a result of reporting and using scores but also have an impact on score meaning and validity (Hubley & Zumbo, 2011). For example, in the case of a depression screen for older adults, consider (a) a potential intended consequence of increased identification of depressive symptomatology in older adults and (b) a potential unintended side effect of greater increases to health insurance rates for older adults. How do each of these social consequences and side effects of use affect the meaning of the depression screening test scores, how depression is conceptualized, and theories about depression, negative affect, and aging? Explicit consideration of social and personal consequences and side effects might enlighten one with respect to whether personal (e.g., age, gender, culture, language) and contextual factors (e.g., poverty, social support, institutionalization) are part of the depression construct or external to it. As Hubley and Zumbo (2011) suggested, when the social consequences and side effects of using a depression screening measure for older adults are not congruent with one's societal values and goals regarding mental health and aging, such insights may be used to modify constructs, theories, and aspects of the test development until the desired congruence between values, purposes, and consequences is accomplished.

CONCLUSION

By design, this chapter has provided an overview of the numerous approaches, models, and foci involved in

investigating the psychometric characteristics of assessment procedures. This chapter's emphasis has been on integrating the sometimes disparate techniques, tools, and test development strategies (including measurement error, reliability, scaling, and score use and reporting) through the lens of validity and validation. We have highlighted the distinctions between observed variable frameworks (i.e., CTT and generalization theory) and latent variable frameworks (i.e., exploratory factor analysis, confirmatory factor analysis, and IRT) because we believe that it is important to understand that the use of a psychometric model is always the choice of the test developer and data analyst and is not necessitated by the data. More often than not, the choice of a particular model is the result of personal beliefs and values, training, and working conventions (Zumbo & Rupp, 2004). Yet the various choices made throughout the measurement and assessment framework have important consequences for the definition, quantification, and use of tests and measures and the decisions that are based thereon. Choosing a psychometric model, or a psychometric technique, is an empirical commitment that demands testing professionals take responsibility for the consequences imparted on the respondents by this choice. As discussed earlier, these demands are present right from the start, wherein the conceptualization and description of constructs and the form of the tests or scales developed to measure them will be heavily influenced by the theories, values, and social context through which they emerge. In this light, as Zumbo (2007) reminded us, it is important to keep in mind that the main goal of investigating the psychometric characteristics of assessment procedures should always be to make valid inferences about the respondents. Working with increasingly more complex and hyperparameterized psychometric models cannot in and of itself increase the validity of these inferences. No matter how sophisticated one's choice of psychometric model, item scoring, item types, and test delivery, scaling, and statistical estimation routines, a poorly conceived and developed test will always remain so.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press. (Original work published 1979)

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397–438. doi:10.1080/10705510903008204

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*, 361–370. doi:10.1177/00131640021970592

Bock, R. D., Mislevy, R. J., & Woodson, C. E. M. (1982). The next stage in educational assessment. *Educational Researcher, 11*, 4–11, 16.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling* (2nd ed.). New York, NY: Springer.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Lanham, MD: Rowman & Littlefield.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0046016

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 465–476. doi:10.1111/j.1754-9434.2010.01273.x

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement, 22*, 15–25.

Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research, 103*, 231–265.

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150. doi:10.2307/2086306

Guttman, L. A. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, & E. A. Schuman (Eds.), *Studies in social psychology in World War II: Vol. 4. Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin, 18,* 98–99.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238–247. doi:10.1037/1040-3590.7.3.238

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123*, 207–215. doi:10.1080/00221309.1996.9921273

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575. doi:10.1111/j.1744-6570.1975.tb01393.x

Likert, R., Roslow, S., & Murphy, G. (1993). A simple and reliable method of scoring the Thurstone Attitude Scales. *Personnel Psychology, 46*, 689–690. doi:10.1111/j.1744-6570.1993.tb00893.x

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382–385. doi:10.1097/00006199-198611000-00017

Martin, R. P. (1988). Basic methods of objective test construction. In R. P. Martin, *Assessment of personality and behavior problems: Infancy through adolescence* (pp. 43–67). New York, NY: Guilford Press.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027. doi:10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*, 35–44. doi:10.1023/A:1006964925094

Ozer, D. J., & Reise, S. P. (1994). Personality assessment. *Annual Review of Psychology, 45*, 357–388. doi:10.1146/annurev.ps.45.020194.002041

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*, 9–13. doi:10.1111/j.1745-3992.1997.tb00586.x

Russell, L. B., & Hubley, A. M. (2005). Importance ratings and weighting: Old concerns and new perspectives. *International Journal of Testing, 5*, 105–130. doi:10.1207/s15327574ijt0502_2

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565–578. doi:10.1037/h0057079

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451. doi:10.1037/h0073357

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research Online, 8*, 1–22.

Wilson, M. (2004). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45–79). Amsterdam, the Netherlands: Elsevier Science.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte, NC: Information Age.

Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bovaird, K. Geisinger, & C. Buckendahl (Eds.), *High stakes testing in education: Science and practice in K–12 settings* (pp. 177–190). Washington, DC: American Psychological Association. doi:10.1037/12330-011

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21–29.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage.

CHAPTER 2

RELIABILITY

Kurt F. Geisinger

Imagine people who have a bathroom scale at home and who as part of their daily morning routine use their bathroom scale to determine their weight. Suppose that on a given morning, a given person weighs in at 165 pounds. One should not consider this individual's true weight to be exactly 165 but rather somewhere within a range, perhaps 164.5 to 165.5 pounds, perhaps 163 to 167, or perhaps even 160 to 170. The width of this range depends on several factors, such as the accuracy of the scale, how well one can see to read the scale, and what clothes the person is wearing when stepping on the scale. The accuracy of such a scale may vary in two ways, and the distinction is an important one. First, it is possible that the scale is consistently high or low. However, this sort of problem is probably of little consequence because one would likely, over time, discover this failing of the scale and adjust the scale accordingly. Errors of this type are called *systematic errors.*

It is also possible that the scale is subject to many other factors that are less obvious—atmospheric pressure, temperature, humidity, where one stands on the scale, and so forth. It is unlikely that even the most weight-conscious of people could correct for these factors in the same way as they could if the scale were always simply 3 pounds too heavy or 5 pounds too light. If weight researchers had the necessary time, energy, and instruments, they might be able to measure and adjust for all of the factors that, taken together, might be considered *random errors.* In other words, random errors are generally not due to any mysterious or mystical forces. Rather, they are considered to be random because people do not know (or do not choose to know) the reasons for their occurrence.

To the extent that the scale is subject to such random errors, it is said that the scale is not reliable. Reliability may be defined and assessed in a number of different ways. For example, it may be defined as the agreement between the bathroom scale in question and some hypothetical, perfect scale; in such a case, reliability might be referred to as accuracy. Because no such perfect scale ever exists, most psychologists use other methods to determine reliability, methods that do not rely on a perfect measuring device. Instead, agreement with other comparable scales, or even with repeated measurements on the same scale, is used. Sometimes, for example, when one does not believe the value that the scale reports, one steps off the scale and tries it again. If the scale is defective, however, it may simply yield the same consistently incorrect measurement on this second occasion (in which case, one may buy a new scale). In such an instance, however, validity rather than reliability is in question, and reliability is generally accepted as a necessary precondition for validity (see Chapter 4, this volume). Similarly, weight-conscious students might also use the scale at home in the morning and then step on one in the nurse's office at school later that same morning. One problem with this latter approach is called breakfast; if these students have eaten between the two weighings, then the two values will not agree; they should not because their weight has in actuality changed from the first measurement to the second. Psychologists use the agreement among seemingly similar

DOI: 10.1037/14047-002
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

and comparable measurement devices to assess the reliability of tests and other forms of measurement, just as this example has used scales. In this chapter, the author first explains the concept of reliability and then describes various methods for assessing the degree to which psychological measures are reliable. Validity is more important than reliability to be sure, but reliability is generally considered to be one of the most important criteria against which measures and, more appropriately, the scores that result from them are evaluated. After a brief section that describes two general models of reliability, some statements regarding the importance of reliability are provided, followed by a discussion of the indices used to evaluate reliability and typical methods used to assess it.

SOME THEORETICAL CONSIDERATIONS

As noted, reliability is considered to be one of the essential characteristics in psychological measurement. The key word that describes reliability is *consistency*. Reliability has as its basis a philosophical notion, what psychometricians have traditionally called *true score*. True scores are hypothetical values that psychologists use to understand and interpret test scores. Psychologists conceive of test scores as being composed of two components: *true score* and *error scores*. For now, think of a true score as the score that an individual would theoretically achieve on a perfectly accurate test, one that was not limited to integer scores. In fact, there are no perfect tests; every existing test is imperfect and therefore provides scores that differ to some extent because of error. This difference between a score earned on the actual test and a score earned on a hypothetically perfect test is referred to as an *error score,* because it indicates the extent to which the observed test score is in error. Measurement error is seen, as in the scale example at the beginning of the chapter, to be randomly occurring. That these errors occur randomly also means that they are assumed to be uncorrelated with true scores or with errors on any other test by the same individuals. These assumptions constitute the basis of what psychometricians have called *classical test theory,* as contrasted with item response theory (see Chapter 6, this volume). Psychometricians now know that the assumption that error is always uncorrelated with other variables is probably not always appropriate (e.g., Green & Hershberger, 2000; Zimmerman, Zumbo, & Lalonde, 1993), as described later in this chapter.

Traditional Sources of Error

Because there are no perfect tests, one can never measure either true or error scores directly, but error variance can be estimated by looking at how a person's score changes over time or differs over comparable tests. When two sets of test scores have been collected (typically, either the same test administered to the same individuals twice or two different forms of the same test administered to the same individuals), the consistency of scores and, conversely, their inconsistency can be estimated. Because true scores are assumed to be invariant, differences among test scores are assumed to be based on differences in error across the two testings. Thus, reliability concerns the extent of agreement between two or more presumably comparable measurement procedures, such as psychological tests, each of which was designed to measure the same variable. This comparability can be defined in numerous ways; basically, each of these ways defines a theoretical position for measuring reliability as well as an operational method of actually estimating the reliability of a set of test scores. Thus, *reliability estimation* refers to a family of procedures rather than a single procedure. The reliability of a set of scores for a specific test is typically expressed as a *reliability coefficient,* an index that enables psychologists to determine whether scores resulting from a test have adequate reliability to be used in a given situation or to compare the reliability of each of several tests. There are different kinds of reliability, all of which use seemingly comparable indices. (Additional attention is provided to reliability coefficients later in this chapter.) Essentially, each theoretical approach leads to a somewhat different kind of reliability coefficient for the scores generated by a given test. For example, *comparability* is sometimes defined as similarity of test scores obtained over time; this form of reliability is sometimes called *stability*. Other times, reliability is seen as similar scores (or at least similar relative positions within

the distribution of test takers) earned by the same people on two alternate forms of the same examination; this approach is called *equivalence*. When the agreement is assessed among a set of comparable measures—most commonly questions composing a single examination—this kind of reliability is called *internal consistency*. The reliability of human judges is actually a special case of internal consistency when there are many judges. A last model, the *generalizability model,* is a rigorous research approach to analyzing consistency (see Chapter 3, this volume). As such, generalizability encompasses all of the previously mentioned reliability models and permits a test developer to design a reliability study tailored to assessing the reliability required by his or her particular needs.

The two variables over which tests are most typically expected to be consistent, if they are to generate reliable scores, are time and content. Measurements of many psychological characteristics are expected to be accurate over a reasonable period of time. If they are not, psychologists may conclude that they are not reliable, or at least not stable, as in the case of psychological states, which are characteristics that are generally changeable. This concept is referred to as *stability* or *temporal consistency*. Similarly, psychometricians expect scores on many psychological measures to generalize to other measures of the construct and to behaviors similar to those specifically measured by the test. For example, suppose an individual has scored highly on a test of intellectual skills. Many intellectual tasks could perhaps have been included on the test but, by virtue of the time limitation of the test administration, were not. It is reasonable to assume that this individual would perform similarly well on these other tasks also. This concept is known as *content sampling.* There are also dimensions other than time and content over which scores on a test should be consistent. Some of these are discussed later in the Generalizability Model section and include the examiner, the testing conditions, and the test format.

Importance of Reliability

That reliability is critical is not always immediately apparent. Two reasons for estimating the reliability of test scores in a given sample may be identified. If research, both pure and applied, is to prove worthwhile, measures must generate scores that are reliable. (It needs to be emphasized that tests are not reliable or unreliable. The scores that such tests produce are what is reliable. Thus, it is possible for a given test to generate reliable scores in one context and not in another.) Also, if credence is to be given to test scores, they must be of reasonable reliability.

Among the goals of research in psychology is the desire to examine the relationships among psychological constructs. Applied psychologists use the functional relationships that emerge from research studies to improve the functioning of people as well as various psychological and social institutions. When variables are not reliably measured, the scores that result are composed largely of random error and do not adequately reflect the underlying true score variable. Thus, the search for statistical indications of consistency among variables is likely to remain fruitless. More troubling is the occasional common occurrence in which the errors of measurement from two unreliable variables correlate with each other in a statistically significant manner. This finding is sometimes called a *local dependence*. These correlations can occur on self-report measures (see Chapter 6, this volume), on ratings among judges who communicate with one another, and on reading test items that are based on the same reading material. When a correlation of this sort is identified, the effects of the local independence violation are generally to overestimate reliability by essentially asking the same question twice. Also, a completely unreliable test can measure nothing because the numerical scores that are assigned to the examinees are essentially random numbers and will not bear a consistent relationship to any other attribute among the individuals being assessed.

Another problem that results from unreliability in psychological tests relates to the interpretation of test scores. Imagine the following fictitious example. A guidance counselor is meeting with Joseph R., a junior in high school who desperately hopes to attend Ivy Halls College, the highly selective institution of higher education that his father attended. The student has just taken the American Scholastic Aptitude Assessment, a test that the

counselor knows Ivy Halls weighs heavily in making admissions decisions. Suppose the counselor knows that Joseph R. has earned scores on the assessment that give him only a 10% chance of admission to Ivy Halls. Should the counselor explore the possibility of attending alternate colleges with Joseph R.? The answer to this question depends in part on the test's reliability. Suppose it yields scores so unreliable that when Joseph takes it a second time, he earns a significantly higher score—one giving him a 60% chance of admission to Ivy Halls College. It is obvious that if the scores generated by this test are relatively unreliable, the counselor would probably be wise to advise Joseph to take the test again (and again, if necessary). However, if the test generates scores that are highly reliable and stable over time, the counselor would probably try to interest Joseph in other colleges or at least to emphasize different aspects of his record in his application. Clearly, the reliability of the test would influence the interpretation that the counselor would make. Imagine the trouble school psychologists would have making decisions about assigning children to special education programs if their tests yielded unreliable scores.

These examples demonstrate the importance of test reliability, both for the advancement of psychological science and for the interpretation of test scores in practical situations. Next, three models used to estimate reliability are discussed.

MODELS OF TEST RELIABILITY

Lord and Novick (1968) presented a number of models of reliability. These models include the parallel testing model and the domain sampling model, both of which are described in this section.

Parallel Testing Model

The parallel testing model, sometimes known as the *classical theory of reliability,* has defined observed test scores and other measurements as consisting of two components: true scores and error scores. To denote the composition of an observed test score (X), the following equation is generally used:

$$X = T + e, \tag{2.1}$$

where X is the observed test score, T is the true score component, and e is the error score component. This theoretical model has been studied empirically and found to be useful in a wide variety of situations over a period of decades, and it continues to be refined. In general, the model proposes that two comparable testings be made and that error variance be estimated from the differences between the two testings. In a strict sense, parallel tests must measure the tested construct equally well. That is, parallel tests would have the same number of items, the same mean, the same variance, and the same true score variance; cover the same constructs or content; have the same correlations with other tests and variables because they measure the same true scores on the parallel tests; and consist of test items that would all measure the construct equivalently. These criteria are difficult to attain in practice. Given that the two tests are parallel, reliability may be investigated by means of one of three methods. (Investigators look at distributions of scores, item intercorrelations, item means, test means and variances, and other information to judge the degree to which two measures are parallel.) Although each of these methods is covered in subsequent sections of this chapter, a brief explanation of each follows. In the first method, known as *test–retest reliability,* a test is administered once, then readministered at a second testing, usually with an intervening period of time. In the second method, known as *alternate-forms reliability,* two versions of a test are developed and administered. Similar test construction practices must be followed rigorously for each form of the test. In the third and final method, a single test is divided into two purportedly equal halves, and each half is said to be parallel to the other. This technique is known as *split-half reliability*.

True scores have been defined in a number of ways in the parallel testing model. For example, one conception of true scores, called *platonic true scores,* portrays true scores as impossible-to-achieve, error-free scores. Thus, if a psychologist were able to look into a person's head and find the actual level of a characteristic in that person, the platonic true score would be found. This portrayal has been criticized by many psychologists (e.g., Thorndike, 1964) in that it implies both that psychological characteristics

exist in concrete form and that the amount of each characteristic does not change over time. Because this idealistic notion of scores is lacking, other conceptions of true scores have also been advanced. More typically, true scores are defined as the average of a very large number of testings; that is, true scores are the expected value of the observed scores over repeated testings of parallel measures. True scores are also defined algebraically in terms of their relationships with observed and error scores.

Adaptations of the parallel testing model have defined true scores by their relationship to error scores and observed scores (Gulliksen, 1950; Lord & Novick, 1968; Nunnally, 1978). For example, it is frequently assumed that error scores are normally distributed over a number of measurements for any given examinee, that true and error scores are uncorrelated, and that error scores are uncorrelated across testings.

Domain Sampling Model

It may be clear from the preceding description of the parallel testing model that it is difficult to develop two truly parallel tests. Even if a test is itself readministered on a second occasion, the examinees may approach this second testing differently, and, hence, the resulting scores and testing may not be truly equivalent. Thus, a more realistic conceptualization of test performance conceives of true scores as being the mean (or average) score if the various ways of measuring the construct were administered an infinite, or at least a large, number of times or if a test taker were assessed by the entire universe of item content. Of course, one aspect of this model is unrealistic: It is difficult to imagine a psychological characteristic that would not be altered by many testings. However, this approach does seem compatible with statistical reasoning (the usefulness of using averages as best guesses of a person's score when no other information is known) and may be used as an idealized perspective on test scores.

The domain sampling model assesses reliability in the following manner. A number of assessments are made. (A domain is specified, such as, e.g., material covered in an introduction to psychology class; then different aspects of that domain are carefully and completely articulated, as in the case of cognitive psychology, social psychology, personality, etc.) Next, items are written to cover these different aspects of the domain, and test items that are representative of the domain are administered. The goal is to estimate the proportion of the domain the individual can answer correctly. These assessments may be tests, observations, ratings, or test items. The assumption is then made that the set of assessments is a random sample of the theoretical domain from which all comparable assessments might have been drawn. How the domain itself is defined depends on the nature of the knowledge, characteristic, or behavior being measured as well as the use to which the test is to be put. One can then estimate how well a person might do on the entire domain or universe of potential assessments by drawing a sample of those elements and applying statistics. The larger the sample is, the better the estimate. Error is conceptualized as leading to the diversity among the estimates and can be estimated using the statistics of sampling distributions. Thus, the domain sampling model involves extrapolation. That is, reliability informs test users about the accuracy present in a sample of observations (e.g., a set of test items) used to estimate scores on the entire universe (or domain) of items.

Generalizability Model

The final model of reliability is one in which the importance of true scores is not as central as in the two preceding models. This model expands the notion of domain as presented in the domain sampling model. The concept of sampling observations from a universe of similar observations holds, but the notion of universe increases. Rather than sampling comparable assessments from the universe, as in the domain sampling model, generalizability calls for the sampling of any observation that is related to the domain of interest. Generalizability studies attempt to determine how much variability each of several factors contributes to similarity of judgments. These factors might include differing examiners, testing conditions, content outlines, times, and so forth. True scores are nominally replaced by what have been called *universe scores;* universe scores are the averages of actual test scores that have been administered over all possible testing conditions. (As noted previously, a more complete

explanation of generalizability is found in Chapter 3, this volume.) The behavior of a number of patients at a psychiatric institution is observed in a number of circumstances, by a number of psychologists, at a number of different times. The statistics of generalizability theory permit one to estimate the size of the psychologist and time effects; such factors are often called *facets*. Thus, researchers can determine how highly one psychologist's judgments agree with the universe scores, how highly judgments agree from one time until another, and so on. All of these observations, although they differ from one another systematically, are members of the universe in that they shed light of one sort or another on the universe score. With the proper research design, the extent to which one can generalize from any kind of observation to any other kind, and from any particular pairing of assessment conditions to the true or overall universe score, can be determined.

In the next sections, both the reliability coefficient and the standard error of measurement are described. After these concepts are defined, common methods of calculating the reliability of a set of scores are presented.

RELIABILITY COEFFICIENTS

Reliability coefficients are indices used to characterize the degree to which test scores are reliable. As previously noted, several methods exist for estimating the reliability of psychological tests. The calculation of reliability coefficients is somewhat different for some of the various methods and, more important, the meaning of the coefficient varies according to the type of reliability involved, as described in the following sections. One aspect of reliability coefficients that is similar across the various methods, however, is the conceptual definition of the coefficient. It may be useful to remember that each test score is generally considered to be composed of true and error elements:

$$X = T + e, \tag{2.2}$$

where X is the observed test score, T is the hypothetical true score, and e is the error score.

When a test is administered to a group of examinees, one can typically calculate only observed scores. But imagine if one could know the true and error scores as well as their observed score sum: One could compute the variance of true scores, error scores, and observed scores and their correlations. Under the assumptions of classical test theory (in which true scores and error scores are uncorrelated), the true score variance and error score variance (typically called *true variance* and *error variance,* respectively) would sum to equal the observed score variance. Thus, observed score variance, as with observed scores themselves, can be partitioned into variance attributable to true scores and variance attributable to error scores. This division is represented in the following equation.

$$s_x^2 = s_t^2 + s_e^2, \tag{2.3}$$

where s_x^2 is the observed variance of the entire test, s_t^2 is the true variance of the test (or the variance of the true scores), and s_e^2 is the error variance of the test scores. It should also be evident that only some of the variability of test scores is due to variability in true scores, unless the reliability of the test is a perfect 1.00.

Reliability coefficients are defined as the proportion of total variance that is true:

$$r_{xx} = \frac{s_t^2}{s_x^2}, \tag{2.4}$$

where r_{xx} is the reliability coefficient, s_t^2 is the true variance, and s_x^2 is the total variance of the actual test scores. One can see that if all the variance of a test is true variance, then the reliability coefficient will equal 1.00; such a circumstance would occur only when there is no error involved. Similarly, if the test variance is totally error variance (and hence, devoid of true variance), then the reliability coefficient will be 0.00. These points represent the maximum and minimum reliability coefficients for any given testing.

As implied by previous discussions, true scores cannot be directly observed. Therefore, the various approaches to reliability estimation attempt to quantify the amount of error variance—the extent of inconsistency—in a given set of test scores. Because true scores do not in theory change, this inconsistency typically serves as a direct representation of error variance. Thus, if test scores show little

stability from measurement to measurement, it is reasonable to conclude that errors of measurement are extensive. Hence, Equation 2.4 follows directly from Equations 2.2 and 2.3 and is more typically the method used to estimate the reliability of test scores:

$$r_{xx} = \frac{s_e^2}{s_x^2}, \tag{2.5}$$

where each term is the same as in Equation 2.2. How error variance is calculated is dependent on the needs of the situation, the limitations of research design, and other considerations. A final thought is that Cronbach and Shavelson (2004) stated that the most important index of reliability is not the reliability coefficient but the standard error of measurement. With the standard error of measurement, one can interpret how variable a person's score is likely to be given the kinds of error affecting test performance.

STANDARD ERROR OF MEASUREMENT

The standard error of measurement is the standard deviation of errors. It is critically used in the interpretation of scores and provides information that relates to how much an individual's observed score is likely to vary around the true score. Thus, one can view a range within which a person's true score is highly likely to fall.

Although reliability coefficients are probably the way most psychologists think of test scores in terms of their consistency, the standard error of measurement is actually a far more useful index when one is interpreting an individual test score. Equation 2.5 shows the formula for computing the standard error of measurement in classical test theory. Classical test theory holds that observed test scores are normally distributed around the true score with a standard deviation equal to the standard error of measurement. In practice, people often apply the standard error of measurement to actual test scores to interpret an individual's range of likely performance.

$$s_e = s_x\sqrt{1 - r_{xx}}. \tag{2.6}$$

Chapter 6 in this volume provides a major advance in the understanding of standard errors of measurement. In classical test theory, there is one standard error of measurement for a test in a given sample. Item response theory acknowledges the differential reliability throughout the range of scores and permits the use of conditional standard errors of measurement throughout the range of scores. That is, for each different score there is a different standard error of measurement.

INFLUENCE OF THE SAMPLE TESTED

The estimation of a test's reliability is dependent on factors such as the amount of time between two testings (in test–retest and alternate forms reliability), the similarity in terms of content and psychological demand characteristics between different elements of the test (in the case of alternate forms, split-half, and other internal consistency approaches), and differences in test administration and scoring (in the case of interrater reliability). In this section, how the sample of individual examinees influences the estimation of the reliability of a set of test scores is discussed. Always remember that the analyses of reliability are sample dependent and not population based (Cronbach & Shavelson, 2004; Thompson & Vacha-Haase, 2000).

The prime characteristic of a sample that affects the reliability coefficient is the variance of the total test scores. "The magnitude of reliability coefficients is dependent on the dispersion of true ability in the group tested. The more heterogeneous the group, the higher r [the reliability coefficient] is likely to be for a given test" (Stanley, 1971, p. 362). This principle can be seen in Equations 2.2 and 2.3, provided earlier in the chapter. Remember from Equation 2.2 that the variance of any set of test scores is portrayed as being the sum of its two components (true variance and error variance) and from Equation 2.3 that the reliability coefficient is the ratio of true variance to total variance. From these two equations, one can anticipate what would happen when the range of ability in a reliability study sample is restricted. Both true and total variances are reduced because of the restriction of range in the sample, but error variance remains just as high as ever because the factors that have led to its existence have not abated. Such restrictions of range occur often in the world, especially when samples have been highly selected. Selected groups occur often in educational institutions,

industry, and hospitals in which acceptance decisions narrow the range of the participants in the research study that involves testing. Whenever groups are formed on the basis of quality in a given characteristic or set of characteristics, then the overall range is typically restricted, but error variance (and hence the standard error of measurement) remains more or less constant, so reliability coefficients are reduced.

METHODS OF ESTIMATING RELIABILITY

In the following discussion, four methods for estimating reliability are provided; three of these are typically used for estimating the reliability of psychological tests of the written or paper-and-pencil variety. These three methods are (a) the test–retest method; (b) the alternate forms method; and (c) various internal consistency methods, including the split-half method. A fourth method that is presented, interrater reliability, is typically used to estimate the agreement between judges rather than between tests.

The test–retest and alternate forms methods of assessing reliability can both be considered examples of the parallel test model of reliability. The split-half method may be considered as an example of the parallel test model as well as of internal consistency methods, which normally tend to be domain sampling in orientation. The test–retest and alternate forms approaches are alike in that both purport to examine the similarity of responses in two testings of the same individuals. In the test–retest method, the same test is administered on two separate occasions. In the alternate forms method, different, comparable forms (of the same test) are used. (The split-half method is, in substance, similar.)

Internal consistency methods are generally representative of the domain sampling model in that they determine the extent to which responses to test questions within a single test permit an investigator to estimate how well examinees would do if they were given all possible questions. Internal consistency methods are invariably used anytime many distinct assessments are made. One of the most common examples of their use is in the estimation of the reliability of tests composed of many individual items, such as multiple-choice tests. Interrater reliability is also a method of estimating reliability from a domain sampling perspective. However, rather than assessing the extent of agreement among paper-and-pencil test items, interrater reliability typically determines the extent of consistency among human judges.

Test–Retest Method

In the test–retest method of estimating test reliability, *reliability* is defined as stability. A test is seen as stable if it yields test scores that are consistent over time. One should note that *stability of scores* is defined as persistent relative standing within the group rather than consistent absolute value of score. Thus, consider a physical education class in which students are timed for the 1-mile run at the beginning of the course in the fall and again at the end of the course in the spring. If the relative standing of students from fall to spring is the same, then the reliability coefficient, defined next, approximates 1.00. This coefficient would even be close to 1.00 if all students decreased their running time by a full minute because their relative standings within the group would be unchanged.

The test–retest method of estimating the reliability of test scores is simple in design and analysis but somewhat more difficult in interpretation. The test is administered to a group, a reasonable period of time typically passes, and the same test is readministered to the same group. The *reliability coefficient* is simply the correlation coefficient between individuals' scores from Time 1 to Time 2. If the correlation coefficient between two testings was .86, then the reliability coefficient would be .86. An interpretation of this coefficient would be that 86% of the test's variance is reflective of variation among true scores, as defined in Equation 2.3. If individuals maintain their same relative positions in the group when retested, the correlation between the two would be high; hence, the reliability coefficient would also be high, even if there is an overall increase or decrease in performance. If similarity between the two sets of relative standings is perfect, a reliability coefficient of 1.00 would result. If, however, test performance is not stable and individuals' scores move about within the group, then a low reliability coefficient will

result. The lowest reliability possible would result if scores from the first testing bear no demonstrable relationship to those from the second testing. In this case, the reliability coefficient would equal 0.00.

The general rule for determining whether test–retest reliability is high or low is that anything that makes the conditions and scoring of the two testings more similar will increase the reliability coefficient. Correspondingly, anything that differentiates the two testings will tend to decrease the coefficient. First, if the test responses for the characteristic being measured are susceptible to being remembered from one testing to the next, then the test–retest method may be inappropriate. Individuals will remember the responses that they gave on the first testing and provide the same response on the second testing. In such cases, the test–retest method probably provides less information than other methods of estimating reliability and provides a spuriously high coefficient. In fact, a reliability of 1.00 might not be an indication of reliability at all if reliability is conceived of as stability of (independent) performance rather than as memory of previous responses. At least four factors affect whether memory is likely to exert a major influence on the apparent reliability of the test. These four factors are (a) the length of time between test and retest, (b) the length of the test, (c) the nature of the test materials, and (d) the nature of the characteristic itself. The length of time between the two testings typically affects the amount remembered and, hence, the reliability coefficient for the test. If the time between testings is short, the resulting test–retest reliability coefficient will probably be inflated. Two rules of thumb should be used when designing and describing a test–retest study. First, the time period between the test and the retest should correspond to the length of time between typical testings and the subsequent behavior that the test attempts to predict. Second, when describing a test–retest study, the author should always provide the length of time between the testings so potential test users may judge whether the time period is appropriate for their prospective test use.

If the length of a test itself is short, it is also more likely that responses from one testing will be remembered until the second testing. If the test materials (items, stimuli, problems, etc.) are distinctive, then they are also more likely to be remembered. This scenario is especially relevant when the test is composed of novel problems that need considerable time or thought to reach an initial solution but once solved are easily remembered, perhaps for life. A final factor that affects the amount of influence that memory holds is simply the nature of the characteristic being measured. For example, responses to multiple-choice items querying students' knowledge of psychology are likely to be remembered. Structured questions on a personality questionnaire are also likely to be recalled. However, physiological measurements are less likely to be influenced by memory. Obviously, if a person's height is measured on one day and again the next, memory of the previous day's height would not influence the second measurement.

A second influence on test–retest reliability coefficients is that of the differential practice effect. Practice effects are well known in psychology. Typically, on many tasks performance improves once a person has some experience in performing the task in question; this is a practice effect. Test taking is not an exception. If the performance of all examinees improves equally, then the correlation between the two testings, and hence the test–retest reliability coefficient, will be appropriate. Unfortunately, in many testing situations, different examinees have varying degrees of test-taking experience that may result in unequal performance increments between testings. In that case, the test–retest reliability may not be completely appropriate. Imagine an employment testing situation in a time of recession. Among the job applicants for a given position are those just out of school. These applicants are experienced in taking paper-and-pencil tests in school and, if they have applied for many jobs, in employment settings as well. Consider now the older applicant, one who may have high mental ability but who has lost his or her job because of the recession. He or she may not have taken paper-and-pencil tests for many years. In a first testing situation, older applicants might be those most likely to experience considerable anxiety. However, one would expect their performance to improve dramatically after some test-taking experience. A test–retest study composed of individuals from both groups—recently graduated and older

applicants—would probably yield an underestimate of reliability. The older individuals' improved performance would change their relative positions in the total group, whereas the younger individuals' performance would remain relatively static. In short, taking the first test in the test–retest design might greatly influence those with minimal test-taking experience while not exerting much effect on those with extensive experience. As noted earlier, in such a circumstance test–retest reliability might be considered an inappropriate method to assess consistency and would certainly underestimate the stability of the test.

The next factor is in keeping with the general principle that anything that makes the two testings more comparable increases the test–retest reliability. Sometimes, the first and second testings are either administered differently or scored under different guidelines. For example, an individual test, whether of intellectual performance or personality, might be administered by a different psychologist who operates with a different style or who uses differing strategies of questioning, and so forth. In a group testing situation, environmental conditions such as noise or heat may also jeopardize the comparability of the two testings. Similarly, the tests may be scored differently from the first testing to the second. Such a problem would be unlikely to occur for an objective-type (e.g., multiple-choice) test but would actually be probable for a less highly structured test such as a projective test, an essay test, or an interview.

It should be obvious to the reader by this point that test–retest reliability has limited usefulness to most psychological characteristics in which memory, practice, mental set, previous experience, and testing conditions affect performance. The reader should similarly see that this method of reliability may be valuable and appropriate in the case of physiological and physical variables and some psychological ones in which memory of testing is not important and there is no interest in whether change occurs over time. However, when the interest is related to a psychological state—a psychological variable that is known to change—then test–retest reliability will likely underestimate the value of a measure while yielding good information about the extent to which the construct changes over time.

Another key concept in determining whether to use a test–retest approach concerns whether the sampling of new content is of interest to the individual seeking reliability information. For example, the sampling of content is important in considering the measurement of a student's level of knowledge on material covered in a particular school unit—typically measured with an educational achievement test. One test represents only a single sampling of all the possible questions that might be asked on the topic covered by the particular examination. Administering the same test under similar conditions at a later time in an effort to assess the test's reliability simply does not indicate how similar a person's performance with another, comparable set of questions would be. When such content sampling is thought to be important, as it is for most psychological characteristics, then a method other than test–retest for estimating reliability would be a preferred strategy. The alternate forms method is one such method.

Alternate Forms Method

The alternate forms method of estimating reliability may be understood when compared and contrasted with the test–retest method. It is similar to the test–retest method in that the same individuals are tested twice, typically on two different occasions, although the intervening time may be shorter than with test–retest studies. For example, a single group of individuals are to take Form A of a given intelligence test on Monday and Form B on Wednesday. The correlation coefficient between the two sets of scores serves as the indication of the test's reliability. This method differs from the test–retest method in two basic ways, however. First, rather than administering the same exact test on two occasions, two different forms of the same test are developed and administered. The two test forms that are used are expected to be parallel as described previously in the Parallel Testing Model section. In the case of the alternate forms method, parallelism is ensured by the use of a detailed outline or set of test specifications on which each test is constructed. Both tests should be comparable while not overlapping in terms of the actual content of the questions presented on the two tests. In actuality, alternate forms may not meet all the criteria specified earlier.

However, they should be based on the same well-defined domain. That is, they should both follow the same outline of test content and should probably use the same mode of measurement (e.g., an essay test could probably not serve as an alternate form for a multiple-choice test and an interview could probably not serve as an alternate form for an intelligence test). That both tests measure the domain equally well is an important assumption that is virtually impossible to demonstrate. (True alternate forms should be able to be equated using the methods described in Chapter 11, this volume.)

A second difference between alternate forms and test–retest methods concerns the time between the two testings. When using the test–retest method with psychological tests, a considerable period of time generally exists between the two testings. When using the alternate forms method, this time interval may be set at the discretion of the test developer; it is not an essential quality of the study. The reliability coefficient in an alternate forms design is the correlation coefficient between scores earned by individuals on the two tests. If the two tests are administered basically back to back—one after the other or on consecutive days—the reliability coefficient is considered to be a coefficient of equivalence; it represents the degree to which the two tests are parallel or equivalent. If the time period between the two tests is extensive enough to resemble the test–retest method, then the reliability coefficient is considered to be a coefficient of stability and equivalence. (Two to 3 weeks is a frequently recommended time interval between the two testings.) Thus, the two factors that are most important in assessing the reliability of scores emerging from a psychological test (sampling of times and sampling of content) are both embodied in this latter method.

A high alternate forms reliability coefficient indicates that the tests appear to be measuring the same underlying true scores. If both tests do a reasonably good job of measuring the true scores, the reliability coefficient, defined as the correlation coefficient between the scores resulting from the two alternate forms, will be high. A low coefficient may indicate one or more of several possibilities. The tests may not measure the same construct. One test form—or both—may simply not measure the construct adequately. Equivalent means, variances, and difficulty levels (in the case of cognitive measures) are also important.

As with the test–retest method, differences in administration of the two test forms will reduce the reliability coefficient. Differences in the scoring of the two forms will similarly reduce the reliability, but one should note that scoring differences may be even more prevalent in the alternate forms method than in the test–retest method because the different set of questions may necessitate different scoring strategies.

The major difference between test–retest and alternate forms is the notion of the content sampling. An example demonstrates this concept. Think of a vocabulary examination given in a foreign language class. The teacher would probably not be able to assess the students on all of the words that they were required to learn. Therefore, the teacher would intend to sample some of them: The teacher might administer a test of 50 English words for which the students must provide the foreign-language equivalents. Although the teacher could administer such an examination on a second occasion and perform a test–retest procedure, the results of the second testing would probably not be seen as a representative sample of the vocabulary domain in that the students had seen them already. To check the alternate forms reliability of the test, one would take a second sample of 50 English words, corresponding to words that the students learned in the foreign language. The teacher might even choose to match the same number of nouns, verbs, adjectives, and so forth on the second test. Thus, the reliability of the test would be seen as how well one test form correlates with another; in this instance, the correlation is between performance on the two test forms. As in the case of test–retest reliability, the correlation coefficient between scores earned on the two measures is the reliability coefficient.

Because alternate forms reliability, as with test–retest reliability, may involve a period of time between the two testings, differences in scoring, and so forth, a psychologist describing an alternate forms study (such as in a manual describing the test) should enumerate the conditions under which the reliability study was performed, with an emphasis on

the amount of time between the two test administrations and any other factors that might have elevated or depressed the correlation. In conclusion, few noncognitive and low-stakes cognitive tests have alternate forms, and not many tests have been subjected to reliability studies of the alternate forms type. It is expensive to design and build an alternate form, especially when one of these forms may not be marketable (because a single test form suffices for many uses). Doubling the costs of examination development simply to provide psychologists with an estimate of reliability may not be defensible in many situations. The expenses involved in building alternate forms of a measure are high enough, and the difficulties inherent in having examinees be examined twice have led to methods for assessing reliability using only a single testing, as described next.

Internal Consistency Methods

Several methods for estimating test reliability are internal consistency methods. Each of these methods relies on estimating the reliability of a test from a content sampling perspective. Unlike the previous two methods, internal consistency methods do not require the administration of two testings; each of these methods looks at the similarity of examinees' responses to various subdivisions of the same test. Such subdivisions may be as large as one half of the entire test or as small as a single test item.

Internal consistency methods share with the alternate forms approach the reliance on content sampling. Whereas the alternate forms approach compares the results of one test form with those of another, internal consistency looks at the agreement among different parts of the same test. When these parts are large (e.g., one half of the test), the internal consistency method closely approaches the alternate forms method. When the parts of the test are small (e.g., individual items), the similarity is still present, although less obvious. Furthermore, internal consistency methods, as with alternate forms reliability, conceptualize reliability primarily in terms of content similarity, but internal consistency methods require only a single testing and therefore do not entail the building, administering, and scoring of two test forms. Thus, this approach allows one to estimate reliability using a single test rather than multiple administrations of identical or alternate forms. The magnitude of such an advantage should be obvious.

The major disadvantage of these methods is that the reliability coefficients solely reflect content sampling and are not sensitive to time sampling. Internal consistency techniques may be divided into two basic types: split-half techniques and item homogeneity techniques. One should note that the split-half techniques preceded the homogeneity techniques historically, and, as one might expect, the split-half methods are considerably less complex computationally than the homogeneity techniques because they were developed largely before the advent of computers and software. The homogeneity approaches subsume the split-half ones; they are essentially more generic approaches to estimating internal consistency reliability than the split-half approaches. Therefore, the distinction between split-half and homogeneity techniques is more apparent than real. Yet the distinction is made for purposes of organization.

Split-half techniques. The simplest of the internal consistency methods are the split-half methods. These methods use the logic of the alternate forms approach by artificially dividing a test into two test halves. The logic of this approach is that if both test halves measure the same quality, examinees should earn comparable scores on each test half; to the extent that such a finding results, the test is seen as being reliable (internally consistent). It is reliable because the two test halves yield comparable information about the person's level of performance. The two test halves operate essentially as parallel forms of a test, although correlating the two halves would only provide the reliability of one half of the test.

Typical approaches for forming the test halves are (a) to match items for content and difficulty on each test half or (b) to divide the test by placing alternate items or test segments into differing test halves. This latter method is most common, especially when the test is artificially divided into odd- and even-item sections for purposes of the reliability analysis; all items are nominally placed into their respective test half for purposes of scoring depending on whether the item number is odd or even.

If items are organized throughout the test, either by difficulty or content, then splitting the test into odd and even halves provides a reasonably easy and equivalent matching of test halves. The correlation coefficient relating the two test halves represents only the reliability of test halves; the coefficient that results is an equivalence coefficient, much as in the case of the alternate forms approach, because it assesses the degree to which the two test halves appear to measure the same quality.

Several statistical methods can be used to compute a reliability coefficient from the two half-test scores. Two methods are presented: the common split-half method and the Guttman method. The most frequent method used is the common split-half method. Table 2.1 demonstrates this method. First, the correlation coefficient between the two (typically, the odd and even) test halves is computed. However, as noted earlier, this correlation coefficient represents the reliability coefficient for each test half rather than for the entire test. An adjustment known as the Spearman–Brown prophecy formula is then applied to the resulting correlation coefficient. The Spearman–Brown formula was developed independently by two researchers to show quantitatively the commonsense notion that the more information one knows about a person, the more reliable judgments about that person will be (Brown, 1910; Spearman, 1910). Operationally, the longer the test is, the more reliable the test data will be. The reliability coefficient calculated in Table 2.1, .892, indicates that almost 90% of the variance of the test results from differences in examinee true scores.

TABLE 2.1

Reliability by the Common Split-Half Method

Individuals	Odd-item score	Even-item score	Total
Vincent	24	20	44
Marie	14	18	32
Jay	19	22	41
Joseph	18	19	37
Gardner	23	24	47
Dorothy	22	24	46
Paul	15	18	33
Anne	24	25	49
Lee	23	25	48
Susanne	16	18	34

Note. Correlation coefficient between odd and even test halves = .805

Spearman-Brown adjustment:

$$r_{xx} = \frac{2r_{oe}}{1+r_{oe}} = \frac{2(.805)}{1+(.805)} = \frac{1.61}{1.805} = .892.$$

Reliability of the entire test = .892.

The general Spearman–Brown formula for estimating the new reliability after a change in test length is

$$r_{nn} = \frac{k(r_{xx})}{1+(k-1)r_{xx}}, \tag{2.7}$$

where r_{nn} is the estimated reliability coefficient after the change in test length, r_{xx} is the reliability coefficient before the change (such as for a test half), and k is the change in test size. The term k is used as follows: If a test has 50 items and the psychologist is anticipating increasing it to 100 items, k would be 2. If the psychologist was considering shortening it to 25 items, then k would be 0.5. The special case of the Spearman–Brown formula for computing the split-half reliability coefficient is easy to determine by substituting 2 for k in Equation 2.6:

$$r_{xx} = \frac{2r_{oe}}{1+r_{oe}}, \tag{2.8}$$

where r_{xx} is the reliability coefficient for the entire test and r_{oe} is the correlation coefficient between the odd and even halves of the test (or any other split of the test into halves). Thus, if two test halves correlate at .50, then when these two halves are combined, the reliability of the entire test would be .67.

This approach to estimating test reliability is relatively easy, and many test manuals report test reliabilities of this type; computation is relatively easy, only one test form is needed, and only one test administration is required to estimate the test's reliability. One problematic assumption of the Spearman–Brown formula is that there are equal variances for each of the halves. When the variances of the two test halves are unequal, the Spearman–Brown formula will yield overestimates of the test reliability (Cronbach, 1951; Gulliksen, 1950). Therefore, other methods to estimate the total test reliability have been developed. Two of these, the Guttman

(1945) and the Rulon (1939) methods, yield identical estimates of the reliability coefficient and are not troubled by differences in the variances of the two halves. The Rulon formula is not presented here; similarly, one may read about the development of the Guttman formula in Guttman (1945) and Haertel (2006). The formula itself is

$$r_{xx} = \frac{4r_{oe} * s_o * s_e}{s_t^2}, \qquad (2.9)$$

where r_{xx} is the total test reliability, r_{oe} is the correlation between the odd and even test halves, s_o is the standard deviation of the odd subtest, s_e is the standard deviation of the even half, and s_t^2 is the variance of the total test.

In Table 2.2, the same data as in Table 2.1 are used to provide an example of the Guttman formula. One should consider the notion of error variance in the Rulon (1939) or Guttman (1945) approaches. Essentially, anything that leads to differences in the scores of examinees from one half to the other contributes to the error variance. Note in Table 2.2 that the variances between the odd- and even-item test halves differ. This difference is what leads to a lower Guttman reliability than is found for the same test data with the common split-half method seen in Table 2.1.

A caution regarding the use of split-half reliability coefficients is in order. Cognitive-type tests differ as being either power tests or speed tests. A true power test is one in which time is not a factor in examinees' performance; basically, there is no time limit, and the examinees have all the time they need to complete the examination. A true speed test is one in which the sole determinant of test performance is how fast an examinee performs. Items are of trivial difficulty if examinees have adequate time. (Tests administered for some clerical positions are true speed tests.) Most educational achievement tests, group intelligence tests, and industrial tests of similar variety are to a greater extent tests of power than tests of speed. Split-half methods of estimating test reliability are not appropriate for tests that are speeded. In a truly speeded test, an individual's score is largely determined by how far he or she gets. Typically, a person answers every question attempted correctly but then misses all items after the final attempted item. Thus, a simple rule is to avoid using split-half formulas when the test in question is highly speeded (i.e., considerably more than just having a time limit). A preferable approach under such circumstances is to use the alternate forms approach or perhaps the test–retest method. Of course, an even simpler reason relates to the fact that these techniques were developed at a time when computations were difficult. Given today's computing power, there is generally no reason to use a split-half approach. Split-half reliability continues to be taught both for historical reasons and to introduce the concept of more general internal consistency approaches; it is unfortunate, therefore, that so many test authors and publishers continue to use this approach to justify the reliability of scores emerging from their measures.

TABLE 2.2

Reliability by Guttman Split-Half Method

Individual	Odd-item score	Even-item score	Total score
Vincent	24	20	44
Marie	14	18	32
Jay	19	22	41
Joseph	18	19	37
Gardner	23	24	47
Dorothy	22	24	46
Paul	15	18	33
Anne	24	25	49
Lee	23	25	48
Susanne	16	18	34
Variance	15.07	9.12	43.66
Standard deviation	3.88	3.02	6.61

Note. Correlation (between odd and even) = .805. Guttman reliability =

$$r_{xx} = \frac{4r_{oe} * s_o * s_e}{s_t^2} = \frac{4 * .805 * 3.88 * 3.02}{43.66} = .864.$$

General internal consistency techniques for test items. It is clear that dividing a test into two seemingly parallel halves is an attempt to establish equivalence within a single test. The question of how to divide a test (e.g., odd vs. even, some content-equated splitting) to establish what might

be considered satisfactory test halves is not easily answered. Reliability coefficients calculated when a test is split in differing ways are likely to differ, sometimes appreciably. How does a test constructor know which to use? And how does a test user know with certainty that the author of the test manual has simply not split the test a number of ways, calculated the various split-half coefficients, and then reported the highest? Also, the frequent failure to meet the common split-half method assumption that the variances of the two test halves be equal (when using psychological tests) explains the development of the Rulon (1939) and Guttman (1945) split-half approaches. It should be obvious that split-half estimates of the reliability of a test, using any of the split-half methods, will typically lead to different estimates of the test's reliability depending on how the test was actually divided (or split) into two halves.

The alternate forms approach to reliability yields a coefficient of equivalence when the two test forms are administered in close temporal proximity. The largest influence on the differences in scores across the two tests is assumed to be the difference in content sampled by the two test forms. Internal consistency methods—split-half and other internal consistency methods—look instead at the degree to which examinee performance is consistent across all the items making up a single test form. The various split-half methods are similar to the alternate forms approach in that they attempt to establish equivalence of test halves by inspecting differences in scores on large segments of the examination. The more general homogeneity approach to internal consistency looks at the comparability of performance across all the items making up a given test. Note that the emphasis is not on the homogeneity of the content, appearance, or format of the questions but on the homogeneity of examinee performance. Two primary factors determine the homogeneity of a test: the number of items on the test and how highly the various items correlate with each other. Thus, the more items on a test, the better the estimate of how well an individual would do if he or she took all items that could possibly be administered; hence, the consistency of the total test scores would be high. Recall that the homogeneity methods of estimating test reliability follow the principles of the domain sampling model. Their goal is to estimate an individual's true score. Simply put, the more appropriate the observations made, the better the estimate of true scores. Hence, as test length increases, true variance increases relative to error variance, with a resultant increase in the reliability coefficient. The other major factor is the interitem correlations; this factor is largely dependent on the similarity of items in terms of their psychological and knowledge demand characteristics. That is, items that call on the same abilities, skills, and experience are likely to be responded to similarly by the individual examinees.

An example may make this point more clear. Suppose there are two tests of introductory French—tests such as would be administered to students in a first course in French. One test could simply be a test of vocabulary; some words are provided in English, and students are expected to provide the corresponding French word, and other words are provided in French, and students give the English translation. The second test might provide the same vocabulary items along with a paragraph to be translated, some questions about French grammar, some questions about proper word use, and perhaps even an oral component in which the student is expected to read a small section of French literature aloud. Although one would expect students who do well on one component of the second test to also do well on others, would the expectation not be that the responses on the former test items would be more similar than those on the latter, given that the content is more similar? That is, it is easier to imagine a student who knows vocabulary and grammar but cannot translate selections or speak French aloud than it is to imagine a student who knows some words and not others in a haphazard manner. Thus, if the domain represented by a test is itself rather heterogeneous, a lower reliability coefficient is likely to result than if the domain were rather tightly defined and each item seemed to measure the same psychological quality.

How broad versus how narrow a construct is also affects the homogeneity of a domain. This factor relates to the narrowness versus the conceptual breadth of the measure and its associated construct. If one is trying to measure something that is relatively narrow (e.g., self-esteem in algebra), item

content is highly similar and item intercorrelations are likely to be quite high. However, a construct such as general self-esteem has very diverse manifestations, and thus the item content is generally quite heterogeneous. In this situation, item intercorrelations are expected to be lower, and thus more items are needed for precise measurement. In fact, if the construct is overly narrow, a researcher could derive a highly precise measurement of a substantive nothing.

The types of reliability coefficients discussed next are all indices or coefficients of homogeneity. Unlike the split-half methods, most yield identical estimates of reliability, given the same test data. Rather, these indices differ in that they are each aimed at different special uses.

Coefficient alpha. The classic method of assessing the reliability of a test in a homogeneity sense is coefficient alpha. Coefficient alpha (Cronbach, 1951) circumvents the problem of how to split a test into two halves because it is mathematically equivalent to the average of all the reliability coefficients that would be computed if the test were split into every possible pair of halves (as calculated using the Guttman [1939] or Rulon [1945] methods). An advantage of using coefficient alpha is that items can be scored on almost any kind of interval scale–level, numerical basis; therefore, this index is often used to assess the reliability of essay tests and attitude scales on which responses are scored, for example, on 1-to-5, 1-to-7, or 1-to-10 bases. Coefficient alpha is generally the procedure of choice when the investigator is interested in the homogeneity among a set of test items making up a test. It is appropriate any time the component parts of a test are summed to form a composite score, as in the case of most educational and psychological tests. The mathematics of the development of coefficient alpha are beyond the scope of this handbook; particulars concerning coefficient alpha may be found in Cronbach (1951), Nunnally (1978), or Cortina (1993). However, the conceptual approach inherent in alpha is simple; the logic behind coefficient alpha is that items that make up a test and that correlate highly among themselves contribute more to the variance of the overall test scores than do items that do not correlate highly with each other. Thus, true variance is that variance that is shared by the items composing the test, the common thread of the test that causes an examinee's performance to be good, medium, or poor. *Error* is defined as anything that influences examinee performance on particular items other than the construct or psychological characteristic underlying the test. Coefficient alpha is calculated as follows:

$$r_{xx} = \frac{k}{k-1}\left[1 - \frac{\sum s_i^2}{s_x^2}\right], \qquad (2.10)$$

where r_{xx} is the reliability coefficient (in this case, coefficient alpha), k is the number of scores being included in the overall test score, $\sum s_i^2$ is the sum of the item variances, and s_x^2 is the variance of the overall test scores. This method of assessing a test's reliability is easy to calculate using a computer and is almost always preferable to the various split-half methods in most instances. Table 2.3 provides an example of the computation of coefficient alpha. The coefficient computed in Table 2.3, .93, indicates that the items appear quite homogeneous. Thus, much of the variation in items is not unique to those specific items and is rather shared among them. Lengthening a test invariably increases the internal consistency reliability and is the most common suggestion for increasing this form of reliability. Another suggestion is to consider item analysis data and remove those items that do not share high correlations with other items or the test as a whole, as is explained in Chapter 7 of this volume. Feldt (1969, 1980) has presented statistical tests that coefficient alpha is equal for test scores emerging from two different tests.

Coefficient alpha is sometimes seen as a measure of unidimensionality, that is, that the test is well represented by a single underlying factor (see Chapter 5, this volume, on the factor analysis of tests and items). This interpretation is generally not accurate (Cortina, 1993; Green, Lissitz, & Mulaik, 1977). Under some circumstances, coefficient alpha can be large even when the test clearly measures several distinct but correlated factors. Because alpha can appear reasonably high even when the measure under consideration is multidimensional (if those dimensions are highly intercorrelated), one must question its use in corrections for attenuation (Schmitt, 1996). Also, when tests are composed of many scales and an overall alpha reliability

TABLE 2.3

Reliability Computation of Coefficient Alpha

Individual	A	B	C	D	E	Total test score
Vincent	10	10	9	9	10	48
Marie	9	9	10	8	9	45
Jay	8	10	10	9	8	45
Joseph	6	5	7	7	8	33
Gardner	10	9	9	8	9	45
Dorothy	8	6	7	8	7	36
Paul	6	6	7	6	7	33
Anne	8	7	7	8	7	37
Lee	9	9	9	10	10	47
Susanne	6	7	7	7	7	34
Variance	2.20	2.96	1.56	1.20	1.36	36.16

Note. These data represent a five-essay final examination. Calculation of coefficient alpha:

$$r_{xx} = \frac{k}{k-1}\left[1 - \frac{\sum s_i^2}{s_x^2}\right] = \frac{5}{4}\left[1 - \frac{2.20+2.96+1.56+1.20+1.36}{36.16}\right] = 1.25\left[1 - \frac{9.28}{36.16}\right] = 1.25(.743) = .93.$$

coefficient is provided, such indices as well as their intercorrelations should also be provided for the individual scales.

Several criticisms of alpha have appeared in recent years (e.g., Bentler, 2009; Green & Yang, 2009; Hattie, 1985; Sitjsma, 2009). Criticisms of coefficient alpha have centered on the just-noted fact that it is a poor index of item unidimensionality and, because of this factor, provides a lower bound estimate of the internal consistency reliability of a set of test scores (Graham, 2006). Structural equation modeling (and factor analyses; see Chapter 5, this volume) has permitted investigations into the structure of tests. In fact, it has become clear that coefficient alpha and most other internal consistency estimates of reliability require a variety of assumptions (Raykov, 1997). When the items composing scores emerging from a measure are not unidimensional, the resultant analyses can lead to inaccurate estimates of internal consistency (Miller, 1995). In this chapter, parallel testing and domain sampling models have been presented; recent research has subdivided these models into tau-equivalent, essential tau-equivalent, and congeneric models, each of which has somewhat different assumptions regarding the degree to which items composing a test measure the underlying construct equivalently, have equal amounts of error variance associated with them, and are essentially differentially precise (Falk & Savalei, 2011; Graham, 2006; Raykov, 1997). Cortina (1993) and Schmitt (1996) attempted to reconcile questions about the appropriateness of coefficient alpha by suggesting that the internal consistency of a set of items includes two considerations: item internal consistency and homogeneity of items. The former relates only to the statistical relationships among items and the latter to unidimensionality; alpha is appropriate for the former but not the latter.

Other internal consistency methods. The commonly used index known as the Kuder–Richardson formula 20, or K–R 20 (Kuder & Richardson, 1937), is a special case formula for coefficient alpha, appropriate when all the items on a test are scored dichotomously—typically either right or wrong. Such items are scored numerically as 0 for wrong responses and 1 for correct responses and are, hence, dichotomously scored. Most objective-type cognitive test items (e.g., multiple choice or true–false) are scored in this way, and an individual's resultant test score equals the sum of all that individual's correct answers. The only difference between the formulae for K–R 20 and coefficient alpha is that the $\sum s_i^2$ term in coefficient alpha becomes $\sum pq$ in the K–R 20;

this change follows because the variance of any dichotomously scored variable is *pq,* where *p* is the proportion of examinees who passed (answered correctly) the item and *q* is the proportion of examinees who fail to answer the item correctly. Hence, the use of *pq* instead of s_i^2 is simply a straightforward substitution.

Interrater reliability. Many kinds of competition involve judgments in scoring: diving, gymnastics, and extreme sports, for example. After each participant performs, the various judges reflect for a moment and then each enters or holds up a score that they believe appropriate for the previous performance. In watching such competition, commentators have wondered aloud how the judges achieved such consistency. In judging such performances, the scoring rules, training, and experience in judging competition is what permits such consistency. More troubling still is when the judges disagree. They may be sitting in different places and their line of sight (visual perspective) makes the performance look different and results in varying evaluations. More troubling would be if their personal perspectives on the performances or national allegiances differ. That such differential perceptions might be held explains the existence of scoring rules. Judgments similar to those made by judges in athletic competitions are made in psychology and education everyday. Teachers read term papers and essays; clinicians determine which diagnosis to make for a client and what kind of treatment to provide. Such decisions demand the analysis of interrater reliability. Therefore, many books concerned with educational and psychological measurement or reliability discuss interrater reliability, which is sometimes called *interrater agreement* or *scorer reliability,* even though it is basically just a form of internal consistency reliability. This use of reliability appears somewhat different from the other methods of estimating reliability, which generally concern objective-type tests. It is an interpretive difference, however, and not a computational one. That is, although this use of the concept of reliability is different, the actual methods or techniques have already been discussed and are typical of other homogeneity techniques. Without exception, these methods simply index how similar different judges, typically human judges, are able to make ratings or judgments.

As noted, until now techniques that are typically used with psychological tests have been presented. Such tests often use multiple-choice or other similar answer formats in that the examinee chooses which of several responses to make. Many of the most interesting kinds of psychological measurement do not use response methods such as multiple choice. Clinical psychologists assessing a client's personality may administer performance-based personality or projective tests—that is, they may make inferences about the individual's personality by showing the individual unstructured stimuli and asking him or her to discuss them; these psychologists are making judgments that may or may not be highly structured. Knowing whether another clinician would make similar judgments of the same client is useful because it would be difficult to have faith in a psychologist who did not agree with his or her colleagues. Such a determination of agreement concerns interrater reliability. If there are two judges, the calculation of the reliability coefficient between the two judges is simply the correlation coefficient between two sets of judgments, which, to be clear, represents only the reliability of a single judge, each set having been made by a judge and all the sets of ratings having been made on the same set of individuals. Again, the correlation coefficient serves as the reliability coefficient. In many instances, however, more than two raters are available to make assessments. In such instances, an intraclass correlation typically serves as the reliability coefficient. One needs a basic understanding of analysis of variance to understand the intraclass correlation; the approach uses a repeated-measures analysis of variance design. The theoretical formula for the intraclass correlation is

$$\sigma^2(b) / [\sigma^2(b) + \sigma^2(w)], \qquad (2.11)$$

where $s^2(w)$ is the pooled variance within subjects and $s^2(b)$ is the variance of the scores between raters.

A calculation of the standard intraclass correlation coefficient is beyond the scope of handbook but may be found in Chapter 3 of this volume (see also Kenny, Kashy, & Cook, 2006; Shrout, 1998; Shrout & Fleiss, 1979). However, note that if one considers

judges as items on a test, a coefficient alpha reliability test can be performed on the judges. The coefficient that resulted would indicate the reliability of their pooled judgments. After all, the ultimate question is how homogeneous the judgments of the various raters are if the judges' scores are aggregated. The description of any assessment procedure that involves expert judgment should include statements regarding the interrater reliability that has been found for the procedure. Similarly, qualifying descriptive information regarding the education, training, and experience of the examiners should be provided. The scoring of many individual tests is reliable when performed under the proper conditions with well-trained scorers, but not under other conditions; therefore, providing evidence of interrater reliability would be valuable for all potential kinds or levels of users of the assessment procedure.

Various techniques have been developed for increasing interrater reliability. The development of procedures for proper test administration and scoring are paramount. Training the scorers in these procedures is also critical. Studies have demonstrated the effectiveness of rater training on interrater reliability, although in many cases they are more effective in reducing systematic errors in ratings rather than so-called random errors. A final important consideration involves the number of raters or observers. Increasing the number of scorers making quick assessments typically increases the reliability to a greater extent than will having each scorer spend more time in making the assessment (Godshalk, Swineford, & Coffman, 1968). This last consideration makes common sense.

A caution is needed regarding a frequent use of interrater reliability in the modern world. This use is in conjunction with the use of videotape equipment and other observational equipment that preserve images and sound digitally. For example, a test development researcher may videotape examiners assessing some youngsters using an experimental intelligence test. Then, to assess the interrater reliability of the instrument, the researcher shows the videotapes of individual testing sessions to other examiners and asks them to score the examination performance that they have seen on the tape. Such a procedure would probably lead to an overestimate of the test's reliability. The reason for this overestimate is that when one administers a test, one makes all the decisions relating to both administration and scoring. When one scores videotaped assessments administered by another examiner, the variability resulting from administration has been removed and only differences due to scoring remain.

When observers are categorizing behavior rather than rating it on a scale, the percentage of agreement among raters is often provided. Such an index is important when one is concerned with the percentage of absolute agreement. However, this index suffers from overestimations when the probability (or base rate) of assignment into the different categories is high. In fact, when the base rate for one or more categories is quite high, the reliability as portrayed by percentage of agreement is an overestimate, and the kappa index is recommended. The kappa index (Fleiss, 1981) adjusts the proportion of cases in that there is agreement for chance agreements.

SUMMARY OF THE APPROACHES TO ESTIMATING RELIABILITY

The various approaches to reliability that have been described in this chapter may be differentiated using several concepts or dimensions. Probably the most important distinction between the various methods relates to what various techniques' primary source of error variance is. That is, the various methods differ largely in terms of the variables that contribute to or define measurement errors. These primary sources of error are shown in Table 2.4. Among the primary sources of error variance are time sampling, content sampling across test forms, content sampling across test halves, content sampling among items, and differences among raters.

Projecting which kind of reliability coefficient will be highest for any given test is difficult. One sometimes hears rules, such as that test–retest reliability coefficients tend to be higher than internal consistency coefficients, but these rules only apply to a given psychological construct or set of constructs. No rule applies across the board. One must simply consider the differences about which a test user must generalize. Then, an honest, professional test developer will perform a reliability study that

TABLE 2.4

Primary Sources of Error Variance Implied by the Varying Approaches to Reliability

Approach to reliability	Source of error variance
Test–retest (coefficient of stability)	Time differences
Alternate form (coefficient of equivalence)	Content differences across test forms
Alternate form (coefficient of stability and equivalence)	Content differences across test forms Time differences
Split-half across test halves	Content differences
Homogeneity (coefficients of internal consistency: coefficient alpha and Kuder–Richardson)	Content differences within the test
Interrater	Rater differences
Generalizability	Any of the above

provides the evidence (e.g., stability, content sampling) appropriate to the potential users.

The various approaches to reliability may also be distinguished on other dimensions. In considering only those appropriate for traditional educational and psychological tests, there are two distinctions: how many test forms are needed and whether one or two test administrations are required. Only the alternate forms approach, whether the second test form is administered after waiting a time interval or not, uses two forms. However, two test administrations are required by the test–retest method, and the alternate forms approach (coefficients of stability and equivalence) is used. In many testing situations in education and psychology, more than one form is already needed, especially in some annual testing programs in the schools and with admissions measures in higher education. Only one test administration is required by all the internal consistency methods (split-half techniques, coefficient alpha, and the Kuder–Richardson formulas) and the alternate forms approach (coefficient of equivalence). Among the internal consistency methods, two additional distinctions may be made: whether the test is simply divided into halves or whether it is studied at the item level, and whether items must be dichotomously scored or not. In regard to the former distinction, the common split-half, Rulon, and Guttman approaches may only be used with test halves; coefficient alpha and the Kuder–Richardson formulas use individual item data. The Kuder–Richardson formulas may only be used when all items are scored as either 0 or 1. All other formulations may be used with items scored in any numerical manner. Again, the recommendation is that split-half approaches no longer need be considered; this author has not encountered a situation in which alpha is not preferable.

CONCLUSION

Reliability has long been one of the premier foci in the evaluation of tests by analyzing the consistency of test scores that result from test administration. Reliability represents a set of procedures that have been quite stable over the past century of psychological testing. In Brennan's (2006) words, "The generic definition of reliability has remained largely intact—namely reliability refers to the consistency of scores across replications of a measurement procedure" (p. 5). Indeed, in recent years, there have been psychometric advances in the understanding of internal consistency reliability with the advent of structural equation modeling. Most of these advances were foreshadowed by Cronbach (1951) himself in his original article describing coefficient alpha. He perceived the need at that time for multiple types of reliability coefficients: alpha, beta, gamma, and so on. However, once he was the primary individual in the development of generalizability theory (Cronbach, Gleser, Rajaratnam, & Nanda, 1972), he did not see the need to explore these other approaches in as much depth (Brennan, 2006; Cronbach, 2004).

Reliability applies to test scores rather than tests per se. Many have recognized, however, that it is easy to fall into the trap of talking about a test as

being reliable or unreliable. When one consistently finds reliable scores emerging from the administration of a given test, it is not surprising that one refers to the test as reliable. Although technically incorrect, it is also common parlance. One should be attentive to portraying reliability as a characteristic of scores rather than of a test. Nevertheless, there are tests that commonly generate reliable scores and those that do not.

In this chapter, the focus has been on the different types of reliability based on the methods used to estimate the reliability of scores (e.g., test–retest, alternate forms, internal consistency). Over the years, this author has heard too many psychologists refer to a test as reliable on the basis of an internal consistency coefficient and then infer that the test scores are stable over time. Just as there are different types of inferences in terms of validity (see Chapter 4, this volume), there are different types of inference related to the various methods to estimate reliability. Researchers and scholars need to be careful about making incorrect inferences. Moreover, split-half reliability, which was discussed at length in this chapter, should be seen primarily as a historical approach to estimating reliability; it was invented at a time when computers and even optical scanning devices did not exist. There is simply no reason to use these procedures today. Having said that, in the preparation of the *Mental Measurements Yearbook* (e.g., Geisinger, Spies, Carlson, & Plake, 2007), the Buros Center for Testing staff has still found that during early years of the 21st century, between 20% and 30% of test manuals provide such evidence.

Reliability will continue to be a primary technique for the evaluation of measures. Its primary roles continue to be as a necessary requirement for the assessment of validity and as information to be used in construct validation studies.

References

Bentler, P. A. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137–143. doi:10.1007/s11336-008-9100-1

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: American Council on Education/ Praeger.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and practice. *Journal of Applied Psychology, 78*, 98–104. doi:10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555

Cronbach, L. J., Gleser, G. C., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measures: Theory of generalizability of scores and profiles*. New York, NY: Wiley.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391–418. doi:10.1177/0013164404266386

Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment, 93*, 445–453. doi:10.1080/00223891.2011.594129

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder–Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363–373. doi:10.1007/BF02289364

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*, 99–105. doi:10.1007/BF02293600

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: Wiley.

Geisinger, K. F., Spies, R. A., Carlson, J. F., & Plake, B. S. (2007). *The seventeenth mental measurements yearbook*. Lincoln, NE: Buros.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1968). *The measurement of writing ability*. New York, NY: College Board.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*, 930–944. doi:10.1177/0013164406288165

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*, 251–270. doi:10.1207/S15328007SEM0702_6

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827–838. doi:10.1177/001316447703700403

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155–167. doi:10.1007/s11336-008-9099-3

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley. doi:10.1037/13240-000

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282. doi:10.1007/BF02288892

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. New York, NY: Addison-Wesley.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspective of classical test theory and structural equation modeling. *Structural Equation Modeling, 2*, 255–273. doi:10.1080/10705519509540013

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184. doi:10.1177/01466216970212006

Rulon, P. (1939). A simplified procedure for estimating the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99–103.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353. doi:10.1037/1040-3590.8.4.350

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*, 301–317. doi:10.1191/096228098672090967

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. doi:10.1037/0033-2909.86.2.420

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. doi:10.1007/s11336-008-9101-0

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Stanley, J. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174–195.

Thorndike, R. L. (1964). Reliability. In *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 23–32). Princeton, NJ: Educational Testing Service.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53*, 33–49. doi:10.1177/0013164493053001003

CHAPTER 3

THE GENERALIZABILITY OF TEST SCORES

Edward W. Wiley, Noreen M. Webb, and Richard J. Shavelson

A company wants to design an instrument to assess specific aspects of anxiety among children whose parents are going through divorce. They design a number of tasks in which a trained counselor could interact with the child in one of a variety of play situations (e.g., drawing pictures, conversing with hand puppets). They realize that a given child's measured anxiety might vary depending on the counselor with whom the child interacts, the play situation presented, or the occasion on which the play situation occurred. How many types of play situations must be given over how many occasions to dependably measure the level of a child's anxiety?

The U.S. Army wants to train its soldiers in how to best interact with individuals from a different culture during field operations. At the end of an extensive training period, trainees participate in several simulations during which they are presented with a situation typical of those encountered in the field, with paid actors posing as members of the other culture. The entire simulation is rated along several dimensions by expert judges. How many judges should rate each session? How many simulations are needed to get a good estimate of a trainee's likely response in an actual field situation?

The two preceding examples involve studies of reliability—that is, the degree to which scores are consistent across multiple conditions (e.g., items, judges, tasks, testing occasions). This chapter details one theory for addressing such question: generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

GENERALIZABILITY: MOTIVATION AND BASIC CONCEPTS

Consider the following example. Contemporary educational policy initiatives are increasingly focused on improving teacher effectiveness—the measurement of which is increasingly contentious. Current initiatives rely heavily on value-added models based on student standardized test scores to measure teachers' contribution to their students' test score gains (Braun, Chudowsky, & Koenig, 2010). Assessments based on teachers' work samples (e.g., lesson plans) represent an alternative measure of classroom teachers' skills. Scoring such assessments is more complex than scoring the multiple-choice items common to standardized assessments, however. Such work samples must be rated, usually against a standard scoring rubric; if raters vary in the stringency with which they score, such variability will affect a teacher's overall assessment score. Furthermore, individual teachers' work samples may vary; for example, three lesson plans submitted by the same teacher would likely receive different scores. Score variability resulting from factors other than differences in teachers' skills is clearly problematic; the aim of reliability analysis is to estimate the degree to which scores are consistent across multiple raters, measurements, and other sources of variability.

Cronbach et al. (1972) developed generalizability (G) theory as a framework for examining the consistency of behavioral measurements (see also Brennan, 2001; Shavelson & Webb, 1991). G theory

DOI: 10.1037/14047-003
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

provides a framework for modeling factors that contribute to score variability. To what degree do the particular aspects of individual tasks or differences in stringency among raters contribute error variability to observed scores? To improve an assessment's reliability, is one better off (a) increasing the number of assessment tasks, (b) having it scored by a greater number of raters, or (c) some combination of the two?

Quoted in a volume in his honor (Snow & Wiley, 1991), Cronbach looked back on G theory as "a tapestry that interweaves ideas from at least two dozen authors, giving the contributions a more significant pattern" (p. 394). Most prominent in the tapestry of G theory is the greater flexibility in modeling measurement reliability than with the reliability methods of one of the interwoven ideas, classical test theory. The general classical test theory model treats individuals' observed scores, X_{pi}, as the sum of two independent components: individuals' true scores, T_p (representing stable or nonrandom individual differences), and measurement error e_{pi}:

$$X_{pi} = T_p + e_{pi}, \tag{3.1}$$

where X_{pi} is individual p's observed score on test item i, T_p is the individual's true score, and e_{pi} is that individual's inconsistent performance or error on test item i.

If individuals' observed scores vary little from one condition to the next, then the magnitude of measurement error must be small, and a given observed score may be taken as a reliable estimate of that person's true score. If, however, individuals' observed scores vary a great deal from one condition to the next, then the magnitude of measurement error must be large, and a given observed score cannot be considered a reliable estimate of that individual's true score. In the former case, reliability is considered high, whereas in the latter reliability is considered low.

Reliability is more formally taken to represent the proportion of observed score variance ($\sigma^2_{X_{pi}}$) resulting from variance among people's true scores ($\sigma^2_{T_p}$) as opposed to variance attributable to other sources ($\sigma^2_{e_{pi}}$). Classical test theory reliability treats measurement error as a single random score component; factors to which this measurement error might be attributed are not individually specified. G theory extends the classical test theory notion of reliability by providing for the modeling of error variation as attributable to multiple systematic sources of score variability (such as differences in rating stringency among raters or variability in task difficulty) as well as those that remain unknown, and by providing for estimation of the magnitude of each.

In G theory, a single behavioral measurement (such as a rating from a performance assessment) is conceived of as a sample from a universe of admissible observations, which consists of all possible observations on an object of measurement (typically a person) that a decision maker considers to be acceptable substitutes for the observation in hand. Each characteristic of the measurement situation (e.g., assessment task, rater, alternative test form, test item, test occasion) is called a *facet*. Individual instances of a facet (such as a particular assessment item or a particular rater, analogous to a factor in analysis of variance [ANOVA] models) are termed *conditions* (analogous to the levels of factors).

To evaluate the dependability of behavioral measurements, a generalizability (G) study is designed to isolate and estimate variation resulting from the object of measurement and as many facets of measurement error as is reasonably and economically feasible. A decision (D) study uses the information provided by the G study to design the best possible application of the measurement for a particular purpose. In planning the D study, the decision maker defines a *universe of generalization,* the set of facets and their conditions to which he or she wants to generalize, and specifies the proposed interpretation of the measurement. The decision maker uses the information from the G study to evaluate the effectiveness of alternative designs for minimizing error and maximizing reliability. In doing so, G theory distinguishes between decisions that concern the relative ordering of individuals (i.e., norm-referenced interpretations of test scores) and those focused on the absolute level of each individual's performance independent of others' performance (i.e., criterion- or domain-referenced interpretations).

In this chapter, we describe the conception and estimation of reliability in the framework of G theory. Throughout the chapter, we demonstrate various concepts using the example of a work-sample

TABLE 3.1

Lesson Plan Pilot Assessment Example: Random-Effects, Crossed Person × Task × Rater Design

Teacher lesson plan	Rater 1				Rater 2			
	Task 1	Task 2	Task 3	Task 4	Task 1	Task 2	Task 3	Task 4
1	$X_{1,1,1}$	$X_{1,2,1}$	$X_{1,3,1}$	$X_{1,4,1}$	$X_{1,1,2}$	$X_{1,2,2}$	$X_{1,3,2}$	$X_{1,4,2}$
2	$X_{2,1,1}$	$X_{2,2,1}$	$X_{2,3,1}$	$X_{2,4,1}$	$X_{2,1,2}$	$X_{2,2,2}$	$X_{2,3,2}$	$X_{2,4,2}$
3	$X_{3,1,1}$	$X_{3,2,1}$	$X_{3,3,1}$	$X_{3,4,1}$	$X_{3,1,2}$	$X_{3,2,2}$	$X_{3,3,2}$	$X_{3,4,2}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
20	$X_{20,1,1}$	$X_{20,2,1}$	$X_{20,3,1}$	$X_{20,4,1}$	$X_{20,1,2}$	$X_{20,2,2}$	$X_{20,3,2}$	$X_{20,4,2}$

assessment of mathematics and science lesson plans completed by preservice teachers as part of their teacher education programs (e.g., Schalock, Schalock, & Ayres, 2006). For example, consider 20 preservice teacher candidates enrolled in a university practicum course who submit as part of their training four separate lesson plans, each of which was scored holistically (on a scale ranging from 1 to 4) by the same two independent raters.[1] This example involves measurement of teacher skills; analogues in psychological measurement are simple to conceive of. For example, Gleser, Green, and Winget (1978) assessed the generalizability of measures of psychological impairment of disaster survivors. In the study, disaster survivors (here analogous to teacher candidates) participated in two independent interviews (here analogous to lesson plans), and each interview was rated along several dimensions by two independent raters (as in the preceding example).

GENERALIZABILITY ANALYSES OF TEACHER LESSON PLAN RATINGS

For our pilot study example, consider a universe of admissible observations consisting of combinations of lesson plans (hereinafter termed *tasks; t*) and raters (*r*). The decision maker (such as a program faculty member) is interested in the performance of teacher candidates (hereinafter termed *persons*) drawn from a particular population. The object of measurement, then, is persons. Person is not a source of error; hence, it is not considered a facet. Assume each teacher candidate's lesson plan was rated holistically by two independent raters. In G theory, one might refer to such a study as having a person × task × rater, or $p \times t \times r$, two-facet crossed random effects design in which all persons perform all tasks and their performance is scored by all raters (see Table 3.1).

Looking ahead, if each teacher's lesson plan scores vary little from one condition (i.e., combination of task and rater) to the next, then each teacher's observed score must be close to his or her true score. Hence, assuming teachers vary in their true scores, score reliability must be high. If, however, each teacher's lesson plan scores vary substantially from one condition to the next, then the correspondence between observed and true scores must be minimal, error must be substantial, and as a consequence reliability must be low. As we will show, providing a framework for estimating score reliability is but one of the benefits of G theory.

Modeling Observed Score Components

In the lesson plan example, each observed lesson plan score (X_{ptr}) can be decomposed into effects

[1]The design described here is simple so as to help us demonstrate basic concepts; in practice, work samples are rated according to a much richer set of criteria (rather than merely given a holistic rating). As the chapter progresses, certain aspects are made complex and thus the example will more closely resemble G studies common in such contexts.

specific to the teacher (person) completing the lesson plan (task), the actual lesson plan (task) being rated, the rater doing the rating, and all combinations of person, task, and rater:

$$\begin{aligned} X_{ptr} &= \mu \text{ (grand mean)} \\ &+ \mu_p - \mu \text{ (person effect)} \\ &+ \mu_t - \mu \text{ (task effect)} \\ &+ \mu_r - \mu \text{ (rater effect)} \\ &+ \mu_{pt} - \mu_p - \mu_t + \mu \text{ (person} \times \text{task effect)} \\ &+ \mu_{pr} - \mu_p - \mu_r + \mu \text{ (person} \times \text{rater effect)} \\ &+ \mu_{tr} - \mu_t - \mu_r + \mu \text{ (task} \times \text{rater effect)} \\ &+ X_{ptr} - \mu_{pt} - \mu_{pr} - \mu_{tr} + \mu_p + \mu_t + \\ &\quad \mu_r - \mu \text{ (residual).} \end{aligned} \tag{3.2}$$

In G theory, μ_p is called the *universe score* (analogous to T_p). As such, μ_p is defined relative to the universe represented by the G study design. One's universe score, then, is defined as the long-run average or expected value (E) of a person's observed score over the universe of admissible observations in the G study (here, all possible combinations of tasks and raters):

$$\mu_p = \mathrm{E}_t \, \mathrm{E}_r \, X_{ptr}. \tag{3.3}$$

Each teacher's universe score is the value we are most interested in estimating—the score that each teacher would receive independent of the particular task and rater combination that was used to generate that score. The population means for task t and rater r are

$$\mu_t \equiv \mathrm{E}_p \, \mathrm{E}_r \, X_{ptr} \tag{3.4}$$

and

$$\mu_r \equiv \mathrm{E}_p \, \mathrm{E}_t \, X_{ptr}. \tag{3.5}$$

These represent the magnitude of the effects particular to task t and rater r, respectively. A positive value for μ_t, for example, would indicate that task t tended to be easier than the average task. The population means for the interaction effects pt, pr, and tr are

$$\mu_{pt} \equiv \mathrm{E}_r \, X_{ptr}, \tag{3.6}$$

$$\mu_{pr} \equiv \mathrm{E}_t \, X_{ptr}, \tag{3.7}$$

and

$$\mu_{tr} \equiv \mathrm{E}_p \, X_{ptr}, \tag{3.8}$$

respectively. The degree to which, say, rater r rates task t more stringently than she or he does the other tasks will show up in the rater × task effect (μ_{tr}; etc.). Finally, the mean over both the population and the universe (the grand mean) is

$$\mu \equiv \mathrm{E}_p \, \mathrm{E}_t \, \mathrm{E}_r X_{ptr}. \tag{3.9}$$

Little attention is typically paid to the grand mean in a G theory setting.

Partitioning Observed Score Variance

As defined in Equation 3.2, each score component, except for the grand mean, has a distribution. The distribution of $\mu_p - \mu$ has a mean of zero and the variance $\mathrm{E}_p(\mu_p - \mu)^2 = \sigma_p^2$. This variance—the universe score variance—represents the degree to which individuals (in our example, teachers) vary in whatever construct is targeted for measurement. Similarly, the variance component specific to tasks (representing variability in observed difficulty across tasks) has a mean of zero and the variance $\mathrm{E}_t(\mu_t - \mu)^2 = \sigma_t^2$, and so forth for the rater component. The person × task component has a mean of zero and the variance $\mathrm{E}_p\mathrm{E}_t(X_{pt} - \mu_p - \mu_t + \mu)^2 = \sigma_{pt}^2$ and represents the score variability attributable to the person × task interaction (pt). This interaction component indicates the degree to which individuals vary in their relative success across tasks (in other words, the degree to which, say, the relative standing of teachers varies across tasks). The variances for the person × rater and task × rater components are defined analogously (each with a mean of zero). Finally, the residual component has a mean of zero and variance $\mathrm{E}_p\mathrm{E}_t\mathrm{E}_r(X_{ptr} - \mu_{pt} - \mu_{pr} - \mu_{tr} + \mu_{pt} + \mu_{pr} + \mu_{tr} - \mu)^2 = \sigma_{ptr,e}^2$, which indicates the person × task × rater interaction (ptr) confounded with any other error that has not been measured (e). The collection of observed scores, X_{ptr}, has a variance $\mathrm{E}_p\mathrm{E}_t\mathrm{E}_r(X_{ptr} - \mu)^2 = \sigma_{X_{ptr}}^2$, which equals the sum of the variance components:

$$\sigma_{X_{ptr}}^2 = \sigma_p^2 + \sigma_t^2 + \sigma_r^2 + \sigma_{pt}^2 + \sigma_{pr}^2 + \sigma_{tr}^2 + \sigma_{ptr,e}^2. \tag{3.10}$$

Via a G study, the variance components in Equation 3.10 are estimated from sample data collected

using the G study design. For the lesson plan example—a random-effects $p \times t \times r$ (person × task × rater) design in which a random sample of n_t lesson plan submissions from each of n_p teachers are rated by each of n_r raters—variance components may be estimated by substituting expected mean squares with their observed analogues (the mean squares from the ANOVA) and solving the set of equations shown in Table 3.2.

Table 3.3 presents variance components estimated from the lesson plan assessment example. The person component reflects systematic variation among teachers' lesson plan ratings. In an ideal case, most score variability would be attributable to differences among teachers, in which case the variance component estimate of σ_p^2 would be large relative to the other variance component estimates (all of which represent sources of error). This is not the case here, however. The variance component of 0.220 for items ($\hat{\sigma}_t^2$), for example, suggests that tasks vary a great deal in their difficulty and that an observed score is quite sensitive to the particular task from which it came. The small variance component for raters, however, suggests that individual raters vary little in the stringency with which they assign scores.

Observed scores are affected by interactions between sources of error as well; magnitudes of these effects are indicated by the variance component estimates for interactions (here, σ_{pt}^2, σ_{pr}^2, and σ_{tr}^2). Take the variance component for the person × task interaction (σ_{pt}^2). In the lesson plan example, σ_{pt}^2 indicates the degree to which individual teachers vary in their responses across specific lesson plans. In other words, σ_{pt}^2 will be larger in cases in which some teachers struggle more than others with

TABLE 3.2

Expected Mean Square Equations for Random-Effects, Multifacet, Crossed $p \times t \times r$ Design

Source of variation	Variance component	Expected mean square (EMS) equation
Persons (*p*)	σ_p^2	$EMS_p = \sigma_{ptr,e}^2 + n_r\sigma_{pt}^2 + n_t\sigma_{pr}^2 + n_t n_r\sigma_p^2$
Tasks (*t*)	σ_t^2	$EMS_t = \sigma_{ptr,e}^2 + n_p\sigma_{tr}^2 + n_r\sigma_{pt}^2 + n_p n_r\sigma_t^2$
Raters (*r*)	σ_r^2	$EMS_r = \sigma_{ptr,e}^2 + n_p\sigma_{tr}^2 + n_t\sigma_{pr}^2 + n_p n_t\sigma_r^2$
pt	σ_{pt}^2	$EMS_{pt} = \sigma_{ptr,e}^2 + n_r\sigma_{pt}^2$
pr	σ_{pr}^2	$EMS_{pt} = \sigma_{ptr,e}^2 + n_t\sigma_{pr}^2$
tr	σ_{tr}^2	$EMS_{tr} = \sigma_{ptr,e}^2 + n_p\sigma_{tr}^2$
ptr,e	$\sigma_{ptr,e}^2$	$EMS_{ptr,e} = \sigma_{ptr,e}^2$

TABLE 3.3

Variance Component Estimates From Lesson Plan Pilot Assessment Example: Random Effects, Crossed $p \times t \times r$ Design

Source of variation	Mean square estimate	Variance component	Estimate	% total variability
Persons (*p*)	1.736	σ_p^2	0.141	16.6
Tasks (*t*)	9.444	σ_t^2	0.220	26.0
Raters (*r*)	0.676	σ_r^2	0.001	0.1
pt	0.524	σ_{pt}^2	0.066	7.8
pr	0.476	σ_{pr}^2	0.021	2.5
tr	0.512	σ_{tr}^2	0.006	0.7
ptr,e	0.392	$\sigma_{ptr,e}^2$	0.392	46.3
Total			0.847	100.0

particular tasks. In this example, the nonnegligible estimate of σ^2_{pt} (0.066, accounting for 7.8% of the total variability) suggests that the relative standing of teachers varies somewhat from task to task. The other two-way interactions appear to contribute little to the variability of observed scores. The large estimate of $\sigma^2_{ptr,e}$ (46.3% of the total variability) reflects the varying relative standing of person across rater–task combinations and other sources of error not systematically incorporated into the G study.

One consequence of estimating variance components by solving the equations in Table 3.2 is that this method may produce negative estimates of variance components. According to Searle (1971), such estimates may arise because of sampling errors or because of model misspecification. Two alternative approaches to dealing with negative variance component estimates that are small in magnitude are (a) to substitute zero for the negative estimate and carry through the zero in other expected mean square equations from the ANOVA (which produces biased estimates; see Cronbach et al., 1972) or (b) to set any negative estimates of variance components to zero but use the negative estimates in expected mean square equations for the other components (Brennan, 2001). When negative variance components are large in magnitude, or when one wishes to avoid the possibility of negative variance components altogether, approaches involve using likelihood methods (maximum likelihood or restricted maximum likelihood) or Bayesian approaches (Box & Tiao, 1973; Fyans, 1977). More detail regarding these methods can be found later in the chapter.

Generalizability and Decision Studies

Prospective teacher candidates may complete lesson plan assignments geared toward a variety of decisions. For example, program faculty may want to rank teachers in order to identify the top 10% for evaluation purposes. Alternatively, faculty may require teachers to achieve a certain score before allowing them to take up a student teaching assignment. Decisions will usually be based on the mean over multiple observations rather than on a single observation. The mean score over a sample of n'_t tasks and n'_r raters, for example, is denoted as X_{pTR} in contrast to a score over a single task and rater, X_{ptr}. A two-facet, crossed D study design in which decisions are to be made on the basis of X_{pTR} is, then, denoted as $p \times T \times R$.

G theory distinguishes between these two types of decisions, the first of which involves relative (i.e., norm-referenced) decisions and the second of which involves absolute (criterion- or domain-referenced) decisions. A *relative decision* concerns the relative ordering of individuals (e.g., norm-referenced interpretations of test scores). In a fully crossed $p \times T \times R$ design, the total error for relative decisions is defined as

$$\delta_{pTR} = (X_{pTR} - \mu_{TR}) - (\mu_p - \mu), \tag{3.11}$$

where $\mu_p = E_T E_R X_{pTR}$ and $\mu_{TR} = E_p X_{pTR}$.

The variance of the errors for relative decisions is

$$\begin{aligned}\sigma^2_\delta &= E_p E_T E_R \delta^2_{pTR} = \sigma^2_{pT} + \sigma^2_{pR} + \sigma^2_{pTR,e} \\ &= \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{ptr,e}}{n'_t n'_r}.\end{aligned} \tag{3.12}$$

In this fully crossed design, all teachers are rated by the same raters on the same lesson plans. Any systematic differences in tasks or raters will affect all teachers and therefore will not change their relative standing. As such, variance components for tasks, raters, and the task × rater interaction are absent from Equation 3.12.

G theory is primarily focused on magnitudes of variance components and measurement error. However, because it originated as a framework for better understanding error variability used for reliability coefficients, it does provide for estimation of a generalizability coefficient ($E\rho^2$) analogous to the classical test theory reliability coefficient[2]:

$$E\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta}. \tag{3.13}$$

Sample estimates of the parameters in Equation 3.13 are used to estimate the generalizability coefficient:

$$E\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \hat{\sigma}^2_\delta}. \tag{3.14}$$

[2]That is, the ratio of universe score variance to the expected observed score variance.

In contrast to the relative decision described here (identification of the top 10% of teacher candidates), an absolute decision (such as the determination of whether a threshold score has been met) focuses on the absolute level of an individual's performance, independent of others' performances (i.e., criterion- or domain-referenced interpretations). For absolute decisions, the error in a random effects $p \times T \times R$ design is defined as

$$\Delta_{pTR} \equiv X_{pTR} - \mu_p \tag{3.15}$$

and the variance of the errors is

$$\begin{aligned}\sigma^2 &= E_p E_T E_R \hat{}^2_{pTR} \\ &= \sigma^2_T + \sigma^2_R + \sigma^2_{pT} + \sigma^2_{pR} + \sigma^2_{TR} + \sigma^2_{pTR,e} \\ &= \frac{\sigma_t}{n'_t} + \frac{\sigma^2_r}{n'_r} + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{tr}}{n'_t n'_r} + \frac{\sigma^2_{ptr,e}}{n'_t n'_r}.\end{aligned} \tag{3.16}$$

With absolute decisions, the main effects of raters and tasks—the strictness of the raters and the difficulty of the tasks—and the interaction between raters and tasks do influence absolute level of performance and so are included in the definition of measurement error. Note also that $\sigma^2 \geq \sigma^2_\delta$.

G theory provides an index of dependability (Kane & Brennan, 1977) for absolute decisions:

$$\Phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2}. \tag{3.17}$$

Table 3.4 gives the results for various D study configurations from the generalizability analyses of the data in Table 3.1. For the design used to collect the data (four tasks, two raters), $E\hat{\rho}^2 = .650$, which falls short of commonly used reliability benchmarks of .70 and .80 for ranking individuals (i.e., for a relative, or norm-referenced, decision). The index of dependability for making absolute (criterion-based) decisions is substantially lower ($\hat{\Phi} = .516$), mainly because of the inclusion of the fairly large $\hat{\sigma}^2_t$ (0.220) in absolute error variability.

For criterion-referenced decisions involving a fixed cut score (λ), it is possible to define a loss function based on the squared distance from that cut score. In such applications, assuming that λ is a constant that is specified a priori, the error of measurement is

$$_{pTR} = (X_{pTR} - \lambda) - (\mu_p - \lambda) = X_{pTR} - \mu_p, \tag{3.18}$$

TABLE 3.4

Generalizability Analysis of Data From Table 3.1 (Random-Effects, Crossed, $p \times t \times r$ Design)

		Decision study configuration			
Source of variation	**Estimated variance component**	$n'_t = 4$, $n'_r = 2$	$n'_t = 6$, $n'_r = 2$	$n'_t = 8$, $n'_r = 2$	$n'_t = 4$, $n'_r = 3$
Estimated variance components					
Persons (*p*)	0.141	0.141	0.141	0.141	0.141
Tasks (*t*)	0.220	0.055	0.037	0.028	0.055
Raters (*r*)	0.001	0.001	0.001	0.001	0.0003
pt	0.066	0.017	0.011	0.008	0.017
pr	0.021	0.011	0.011	0.011	0.007
tr	0.006	0.001	0.001	0.000	0.001
ptr,e	0.392	0.049	0.033	0.025	0.033
Error variances					
$\hat{\sigma}_\delta$		0.076	0.054	0.043	0.056
$\hat{\sigma}_\Delta$		0.132	0.092	0.072	0.112
Coefficients					
$E\hat{\rho}^2$		.650	.722	.765	.715
$\hat{\Phi}$		.516	.606	.663	.557

and an index of dependability may be defined as

$$\Phi_\lambda = \frac{E_p(\mu_p - \lambda)^2}{E_T E_R E_p(X_{pI} - \lambda)^2} = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma^2}. \quad (3.19)$$

An unbiased estimator of $(\mu - \lambda)^2$ is $(\overline{X} - \lambda)^2 - \hat{\sigma}^2_{\overline{X}}$, where $\overline{X}$ is the observed grand mean over all sampled persons, tasks, and raters and $\hat{\sigma}^2_{\overline{X}}$ is the error variance involved in using $\overline{X}$ as an estimate of μ, the population grand mean (over all persons, items, and raters).

For the $p \times T \times R$ random-effects design, $\hat{\sigma}^2_{\overline{X}}$ is

$$\hat{\sigma}^2_{\overline{X}} = \frac{\hat{\sigma}^2_p}{n'_p} + \frac{\hat{\sigma}^2_t}{n'_t} + \frac{\hat{\sigma}^2_r}{n'_r} + \frac{\hat{\sigma}^2_{pt}}{n'_p n'_t} + \frac{\hat{\sigma}^2_{pr}}{n'_p n'_r} + \frac{\hat{\sigma}^2_{tr}}{n'_t n'_r} + \frac{\hat{\sigma}^2_{ptr,e}}{n'_p n'_t n'_r}. \quad (3.20)$$

The estimate of Φ_λ is smallest when the cut score λ is equal to the observed grand mean $\overline{X}$; in that case, $\hat{\Phi}_\lambda = \hat{\Phi}$.

To this point, we have talked about G studies for assessing magnitudes of sources of sampling variability, for the purpose of assessing reliability for decisions made on the basis of a particular design. G theory also provides for the estimation of reliability under alternative D study designs (i.e., designs other than those used to collect the original data). For instance, in our example, we may want to gauge the effect of doubling our number of tasks from four to eight or of doubling the number of raters from two to four.

The results in Table 3.4 suggest substantial improvements in $E\hat{\rho}^2$ and $\hat{\Phi}$, with increases in numbers of tasks sampled; this is to be expected because $\hat{\sigma}^2_t$ was large in magnitude. Estimated values of generalizability and dependability for six tasks are $E\hat{\rho}^2 = .722$ and $\hat{\Phi} = .606$. Adding another two tasks (for a total of eight) increases generalizability and dependability even more ($E\hat{\rho}^2 = .765$ and $\hat{\Phi} = .663$). Increasing the number of raters has less effect on estimated values of generalizability and dependability. For example, a total of 12 scores per person using four tasks and three raters yields somewhat lower estimates ($E\hat{\rho}^2 = .715$ and $\hat{\Phi} = .557$) than a total of 12 scores per person using six tasks and two raters.

Crossed and Nested Designs

Project constraints often prohibit fully crossed D study designs. In our example, the decision maker may want to sample a larger number of lesson plans for each teacher (e.g., eight instead of four) but do so without increasing the number of raters (staying with two raters) and without increasing the burden for each rater. This can be accomplished by giving each rater responsibility for a different set of four lesson plans. Alternatively, rather than having each rater rate all four lesson plans for every person (a total of 80 plans for each rater), it may be more feasible to use eight raters and have each pair of raters rate a different lesson plan (i.e., each rater rates a total of 20 plans). G theory accommodates nested designs such as these for which not all levels are crossed with all levels of the other facets.

The variance components estimated from a two-facet crossed G study can be used to estimate error variance and generalizability and phi coefficients for a wide range of D study designs, including nested designs. In fact, any G study can be used to estimate the effects in a D study design with the same or more nesting than the G study design. Table 3.5 lists the possible two-facet G and D study designs for which p is the object of measurement and is not nested within a facet.

Corresponding to the first scenario described, a decision maker who uses a crossed $p \times t \times r$ G study design may choose to use a $p \times (T : R)$ design in the D study. This D study design would gauge the effect on reliability if each rater rated a different subset of the tasks. Using the variance components from the $p \times t \times r$ G study design, error variances for relative and absolute decisions for a $p \times (T : R)$ D study design are

$$\sigma^2_\delta = \sigma^2_{pR} + \sigma^2_{pT:R} = \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{pt,ptr,e}}{n'_t n'_r}, \quad (3.21)$$

and

$$\sigma^2 = \sigma^2_R + \sigma^2_{pR} + \sigma^2_{T:R} + \sigma^2_{pT:R} = \frac{\sigma^2_r}{n'_r} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{t,tr}}{n'_t n'_r} + \frac{\sigma^2_{pt,ptr,e}}{n'_t n'_r}, \quad (3.22)$$

TABLE 3.5

Possible Random Effects Decision Study Designs From Random-Effects, Two-Facet, Generalizability Study Design

Generalizability study design	Decision study design
$p \times t \times r$	$p \times T \times R$
	$p \times (T{:}\ R)$ or $p \times (R{:}\ T)$
	$(R{:}\ p) \times T$ or $(T{:}\ p) \times R$
	$R{:}\ (p \times T)$ or $T{:}\ (R \times p)$
	$(R \times T){:}\ p$
	$R{:}\ T{:}\ p$ or $T{:}\ R{:}\ p$
$p \times (r{:}\ t)$	$p \times (R{:}\ T)$
	$R{:}\ T{:}\ p$
$p \times (t{:}\ r)$	$p \times (T{:}\ R)$
	$T{:}\ R{:}\ p$
$r{:}(p \times t)$	$R{:}(p \times T)$
	R: T: p
$t{:}(p \times r)$	$T{:}(p \times R)$
	$T{:}\ R{:}\ p$
$(r \times t){:}\ p$	$(R \times T){:}\ p$
	$R{:}\ T{:}\ p$ or $T{:}\ R{:}\ p$

Note. Lowercase letters refer to facets of the G study, whereas uppercase letters refer to facets of the D study.

where $\sigma^2_{t,tr}$ is the sum of σ^2_t and σ^2_{tr} from the $p \times t \times r$ G study and $\sigma^2_{pr,ptr,e}$ is the sum of σ^2_{pt} and $\sigma^2_{ptr,e}$. For the same number of total observations per teacher (e.g., eight), nesting tasks within raters will result in improved precision compared with a fully crossed design because a larger number of tasks are sampled. Thus, the variance components σ^2_t and σ^2_{pt} are divided by $n'_t n'_r$ (8) instead of n'_t (4).

With values from Table 3.3, Equations 3.21 and 3.22 are used to find error variances for relative and absolute decisions:

$$\hat{\sigma}^2_\delta = \frac{\hat{\sigma}^2_{pr}}{n'_r} + \frac{\hat{\sigma}^2_{pt,ptr,e}}{n'_t n'_r}$$
$$= \frac{0.021}{2} + \frac{(0.066 + 0.392)}{2 * 4}$$
$$= 0.068$$

and

$$\hat{\sigma}^2_\Delta = \frac{\hat{\sigma}^2_r}{n'_r} + \frac{\hat{\sigma}^2_{t,tr}}{n'_t n'_r} + \frac{\hat{\sigma}^2_{pr}}{n'_r} + \frac{\hat{\sigma}^2_{pt,ptr,e}}{n'_t n'_r}$$
$$= \frac{0.001}{2} + \frac{(0.220 + 0.006)}{2 * 4} + \frac{0.021}{2} + \frac{(0.066 + 0.392)}{2 * 4}$$
$$= 0.097$$

Compared with $\hat{\sigma}^2_\delta = 0.076$ and $\hat{\sigma}^2_\Delta = 0.132$ from the original fully crossed G study, the error variances from the nested study are somewhat smaller, and consequently the estimated reliability and dependability coefficients are somewhat larger ($E\hat{\rho}^2 = .675$ and $\hat{\Phi} = .594$ for the nested design compared with $E\hat{\rho}^2 = .650$ and $\hat{\Phi} = .516$ for the crossed design).

RANDOM AND FIXED FACETS

G theory is a theory primarily concerned with random effects. In a G study, facets are typically considered random if conditions of the facet (such as a given lesson plan task) can be said to represent a random sample of all possible conditions of the facet that could have been sampled. The sampling of conditions of a facet from a broader universe is what drives the purpose of invoking G theory. Gauging the variability introduced by this sampling is the primary purpose of a G study. When the levels of a facet have not been sampled randomly from the universe of admissible observations but the intended universe of generalization is infinitely large, the concept of exchangeability may be invoked to consider the facet as random (de Finetti, 1937). Even if conditions of a facet have not been sampled randomly, the facet may be considered to be random if conditions not observed in the G study are exchangeable with the observed conditions.

In contrast to treating a facet as random, a facet is treated as fixed if conditions of that facet have not been sampled from a broader universe (to which the score is to be generalized). Often, a researcher will purposely select a few specific conditions and have neither basis for nor interest in generalizing beyond them; alternatively, the universe of conditions may be small enough to include all possible conditions in the G study. When it makes sense to do so (e.g., when the researcher is interested in a total score across the multiple conditions rather than scores

particular to each condition), the researcher examines average scores across the fixed facets (Brennan, 2001; Cronbach et al., 1972). However, when the researcher wants to maintain the distinction between the conditions of a fixed facet, she or he may either (a) carry out separate G studies within each condition (Shavelson & Webb, 1991) or (b) model data using a multivariate generalizability model that treats each condition of the fixed facet as a separate dimension (Brennan, 2001).

Earlier, we discussed G theory's ability to provide insight into score variability, variance components, and reliability under designs other than that of the G study used to collect the original data (via the D study). (Recall, for example, the many D study designs possible for various G study designs as listed in Table 3.5.) Similarly, through the D study the researcher may use variance components for effects specified as random in the G study to estimate the reliability of scores under the assumption of those same effects as fixed.

Recall the $p \times t \times r$ G study of the lesson plan assessment in which both tasks and raters were modeled as random. If the particular tasks used in the assessment are going to remain constant across subsequent administrations of the assessment and would therefore not be treated as exchangeable with others not under consideration, then the assessor may ultimately decide it is more appropriate to model tasks as fixed rather than random. Such a model is referred to as *mixed* because it includes both random facets (raters) and fixed facets (tasks). If, in such a case, one was interested in a total score across the tasks, we would approach the D study by modeling (for each person–rater combination) the average score over the four tasks as if it was the observed score.

Brennan (2001) details procedures for variance component estimation in mixed-model D studies. Person-level variability, denoted as σ^2_τ to distinguish universe score variance in a mixed model from that in a fully random model, will be affected not only by the person-level variability but also by the interaction between persons and tasks, averaged over the number of conditions of the fixed facet (four tasks in our example):

$$\sigma^2_\tau = \sigma^2_p + \sigma^2_{pT} = \sigma^2_p + \frac{\sigma^2_{pt}}{n'_t}. \tag{3.23}$$

To this point, universe score variance has generally consisted of only σ^2_p. However, the treatment of tasks as fixed—and the average scores over conditions as analogous to the observed scores—introduces the need to generalize the notion of universe score as the expectation over all conditions (e.g., possible tasks) to allow us to model it as a person's expected score over only those conditions of the fixed facet.

In the lesson plan example with tasks (t) fixed, the relative error variance and associated generalizability component are as follows:

$$\sigma^2_\delta = \sigma^2_{pR} + \sigma^2_{pRT} = \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{prt,e}}{n'_r n'_t} \tag{3.24}$$

and

$$E\rho^2 = \frac{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t}}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{prt,e}}{n'_r n'_t}}. \tag{3.25}$$

Similarly, the absolute error variance and index of dependability are

$$\begin{aligned}\sigma^2 &= \sigma^2_R + \sigma^2_{pR} + \sigma^2_{RT} + \sigma^2_{pRT} \\ &= \frac{\sigma^2_r}{n'_r} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{rt}}{n'_r n'_t} + \frac{\sigma^2_{prt,e}}{n'_r n'_t}\end{aligned} \tag{3.26}$$

and

$$\Phi = \frac{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t}}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_r}{n'_r} + \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{rt}}{n'_r n'_t} + \frac{\sigma^2_{prt,e}}{n'_r n'_t}}. \tag{3.27}$$

Consider the consequences of modeling tasks as fixed rather than random in the lesson plan example. Recall that our original G study specified all facets as random ($p \times t \times r$). Using that design, Table 3.4 showed that the fully random design would require upward of eight tasks to approach the typical benchmark of .80 in our reliability-like indices. Table 3.6 provides alternative mixed-effect D study designs in which tasks are specified as fixed.

TABLE 3.6

Lesson Plan Example: Variances and Reliability-Like Coefficients for $p \times t \times r$ Design With Tasks (t) Fixed (Based on Variance Component Estimates From Table 3.3)

	Tasks fixed; designs average over four specific tasks		
Variance	$n'_r = 1$	$n'_r = 2$	$n'_r = 3$
Universe score variance			
$\hat{\sigma}^2_\tau$	0.156	0.156	0.156
Error variances			
$\hat{\sigma}^2_\delta$	0.119	0.060	0.040
$\hat{\sigma}^2_\Delta$	0.122	0.061	0.041
Coefficients			
$\mathbf{E}\hat{\rho}^2$	0.570	0.726	0.799
$\hat{\Phi}$	0.565	0.722	0.795

Reliability-like coefficients are higher when variability resulting from task sampling is excluded from measurement error; for example, the $p \times t \times r$ design with four tasks and two raters yields $\mathrm{E}\hat{\rho}^2 = .650$ and $\hat{\Phi} = .516$ with tasks specified as random, compared with $\mathrm{E}\hat{\rho}^2 = .726$ and $\hat{\Phi} = .722$ when tasks are modeled as fixed.

As mentioned earlier, in some cases it may not be advisable to average over all levels of a fixed facet. If tasks measure quite different constructs, for example (say, mathematics and science skills), an average score across the two would obscure relative performances on the two for any given individual. In this case, the decision maker may wish to investigate dependability for each construct separately. Substantive arguments such as these aside, the G study itself provides quantitative information to help guide a decision of whether to model an effect as fixed or random. For example, a large variance component for an interaction between person and a particular facet (such as $\hat{\sigma}^2_{pt}$) suggests that individuals varied in their performance on the various tasks, which may indicate that the tasks measured different constructs or were eliciting responses that differed across people for some other reason. In any event, in such cases one would want to retain the score information specific to each condition (subject matter, in this case), so one would either (a) carry out separate G studies for each condition or (b) treat each condition as a separate dimension in a multivariate G study model (for details, see Brennan, 2001).

MULTIVARIATE GENERALIZABILITY THEORY

In many cases, researchers are interested in measurement along multiple dimensions, as opposed to a single dimension, as we have done to this point. That is, so far our example has considered only a single, holistic rating for each lesson plan. As noted earlier, however, we could rate each lesson plan according to a much more comprehensive set of criteria (timing of lesson activities, appropriateness of activities, quality of assessments, etc.). In such cases, the G study researcher turns to multivariate G theory. Because of space limitations, we cannot provide a full treatment of multivariate G theory here; for that, we point the reader to Brennan (2001). Here we provide an example of how a multivariate treatment of the teacher lesson plan example might differ from its analogous univariate treatment.

We focus on using multivariate G theory to estimate the generalizability of composite scores (for other uses of multivariate G theory, see Shavelson & Webb, 1981; Webb, Shavelson, & Haertel, 2007). For this example, rather than treating the four lesson plans completed by each teacher as interchangeable, we assume that two of these represent mathematics lesson plans and two of them represent science lesson plans. To reflect specific dimensions in our multivariate study, we denote observed scores on math and science as $_m(X_{ptr})$ and $_s(X_{ptr})$, respectively. Variability of these observed scores is represented by $\sigma^2_{m_{(X_{ptr})}}$ and $\sigma^2_{s_{(X_{ptr})}}$.

In multivariate G theory, the individual variance components of univariate theory generalize to individual covariance matrices. For example, σ^2_p generalizes to

$$\begin{bmatrix} \sigma^2_{mp} & \sigma_{mp,sp} \\ \sigma_{sp,mp} & \sigma^2_{sp} \end{bmatrix}, \tag{3.28}$$

with σ^2_{mp} representing person-level variability in mathematics, σ^2_{sp} representing person-level

variability in science, and $\sigma_{s^p,m^p} = \sigma_{m^p,s^p}$ representing person-level covariance between mathematics and science. We refer to the elements of Equation 3.28 as person-level covariance components.

Recall from Equation 3.10 that in univariate G theory, total score variability $\sigma^2_{X_{ptr}}$ was decomposed into seven independent variance components:

$$\sigma^2_{X_{ptr}} = \sigma^2_p + \sigma^2_t + \sigma^2_r + \sigma^2_{pt} + \sigma^2_{pr} + \sigma^2_{tr} + \sigma^2_{ptr,e}.$$

A parallel decomposition holds for multivariate G theory. Consider universe score variability; with two separate dimensions, universe score variability must include variability in one dimension (mathematics), variability in the other dimension (science), and covariability (i.e., covariance) between the two, as shown in Equation 3.28. To accommodate this, each variance component included in Equation 3.10 generalizes to a 2 × 2 covariance matrix (representing the covariances between the two dimensions):

$$\begin{bmatrix} \sigma^2_{m^{(Xptr)}} & \sigma_{m^{(Xptr)},s^{(Xptr)}} \\ \sigma_{s^{(Xptr)},m^{(Xptr)}} & \sigma^2_{s^{(Xptr)}} \end{bmatrix} =$$
$$\begin{bmatrix} \sigma^2_{m^p} & \sigma_{m^p,s^p} \\ \sigma_{s^p,m^p} & \sigma^2_{s^p} \end{bmatrix} (\text{person}) + \begin{bmatrix} \sigma^2_{m^t} & \sigma_{m^t,s^t} \\ \sigma_{s^t,m^t} & \sigma^2_{s^t} \end{bmatrix} (\text{task})$$
$$+ \begin{bmatrix} \sigma^2_{m^r} & \sigma_{m^r,s^r} \\ \sigma_{s^r,m^r} & \sigma^2_{s^r} \end{bmatrix} (\text{rater})$$
$$+ \begin{bmatrix} \sigma^2_{m^{pt}} & \sigma_{m^{pt},s^{pt}} \\ \sigma_{s^{pt},m^{pt}} & \sigma^2_{s^{pt}} \end{bmatrix} (\text{person} * \text{task})$$
$$+ \begin{bmatrix} \sigma^2_{m^{pr}} & \sigma_{m^{pr},s^{pr}} \\ \sigma_{s^{pr},m^{pr}} & \sigma^2_{s^{pr}} \end{bmatrix} (\text{person} * \text{rater})$$
$$+ \begin{bmatrix} \sigma^2_{m^{tr}} & \sigma_{m^{tr},s^{tr}} \\ \sigma_{s^{tr},m^{tr}} & \sigma^2_{s^{tr}} \end{bmatrix} (\text{task} * \text{rater})$$
$$+ \begin{bmatrix} \sigma^2_{m^{(ptr,e)}} & \sigma_{m^{(ptr,e)},s^{(ptr,e)}} \\ \sigma_{s^{(ptr,e)},m^{(ptr,e)}} & \sigma^2_{s^{(ptr,e)}} \end{bmatrix} (\text{residual}). \quad (3.29$$

In univariate G theory, we had little need to specify covariance components, because all effects were specified to be independent. This is no longer the case in the multivariate form of G theory. Consider, for example, that the same two raters rate both the mathematics and the science lesson plans. In such a case, we might expect rater stringency to carry across the two assessments (e.g., tough raters rate stringently on both mathematics and science assessments). In such a case, we would expect a nonzero (here, positive) correlation between the rater effects in math and rater effects in science. In other words, we would expect $\sigma_{m^r,s^r} \neq 0$.

When expected values of covariance components are nonzero, the two conditions are referred to as *linked* (Cronbach et al., 1972; see also Chapters 10 and 11, this volume)—in other words, observations from the two dimensions share the same conditions. *Unlinked* conditions are those for which the expected value of their covariance is zero; this would have been the case if raters of mathematics lesson plans had been selected from a different rater pool than those who rated science lesson plans.

Joe and Woodward (1976) provided a coefficient of generalizability for the two-facet, crossed, balanced multivariate composite:

$$E\hat{\rho} = \frac{\mathbf{a}'\mathbf{V}_\mathbf{p}\mathbf{a}}{\mathbf{a}'\mathbf{V}_\mathbf{p}\mathbf{a} + \frac{\mathbf{a}'\mathbf{V}_\mathbf{pt}\mathbf{a}}{n'_t} + \frac{\mathbf{a}'\mathbf{V}_\mathbf{pr}\mathbf{a}}{n'_t} + \frac{\mathbf{a}'\mathbf{V}_\mathbf{ptr,e}\mathbf{a}}{(n'_t n'_r)}}, \quad (3.30)$$

where $\mathbf{V}$ is a matrix of variance of variance components and covariance components, and $\mathbf{a}$ is the vector of weights that maximizes the ratio of between-person to between-person plus within-person variance component matrices. Alternatives to maximizing the reliability of a composite are to determine variable weights on the basis of expert judgment or use weights derived from a confirmatory factor analysis (Marcoulides, 1994).

Table 3.7 presents covariance and variance components from the lesson plan example. The values on the diagonal represent variance components specific to a given subject matter. For instance, we may note that universe score variability is slightly greater for science $(\hat{\sigma}^2_{s^p} = 0.162)$ than for mathematics $(\hat{\sigma}^2_{m^p} = 0.120)$. In contrast, person × task variance appears substantial for the mathematics assessment $(\hat{\sigma}^2_{m^{pt}} = 0.120)$; individual teachers varied in the degree to which they

TABLE 3.7

Estimated Variance and Covariance Components of Multivariate Generalizability Study of Teacher Lesson Plan Example

Source of variation	Estimated univariate variance components (facets random)	Estimated covariance and variance components from multivariate generalizability study
		Math Science
Persons (*p*)	0.141	Math [0.120 0.133] Science [0.133 0.162]
Tasks (*t*)	0.220	Math [0.258 (0.123)] Science [(0.123) 0.182]
Raters (*r*)	0.001	Math [0.001 0.0004] Science [0.0004 0.001]
pt	0.066	Math [0.120 (0.005)] Science [(0.005) 0.012]
pr	0.021	Math [0.015 (0.015)] Science [(0.015) 0.027]
tr	0.006	Math [0.004 0.002] Science [0.002 0.004]
ptr, e	0.392	Math [0.374 (0.023)] Science [(0.023) 0.410]
Multivariate generalizability coefficients		
$E\hat{\rho}^2 \left(\frac{1}{2}\text{Math}; \frac{1}{2}\text{Science}\right)$		0.817
$E\hat{\rho}^2 \left(\frac{2}{3}\text{Math}; \frac{1}{3}\text{Science}\right)$		0.775
$E\hat{\rho}^2 \left(\frac{1}{3}\text{Math}; \frac{2}{3}\text{Science}\right)$		0.816

found specific mathematics tasks more difficult than others. The same was not true for science because the corresponding variance component appeared to be very small ($\hat{\sigma}^2_{m^{pt}} = 0.012$).

The off-diagonal elements represent covariance components. The large positive value for the person covariance component ($\hat{\sigma}_{s^p,m^p} = 0.103$) suggests a strong relationship between individuals' ratings on math lesson plans and their ratings on science lesson plans. However, the large negative value for the task covariance component ($\hat{\sigma}_{s^t,n^t_t} = -0.123$) suggests that the tasks that teachers found most difficult for the science lesson plan were relatively easy for the math lesson plan (and vice versa). The relatively small covariance components for the other facets suggest that variability present in one subject was unrelated to variability in the other subject.

The researcher may wish to combine information from each subject matter to create a composite score, the reliability of which can be determined as in Equation 3.30. Doing so requires the specification $\mathbf{a}$, a vector of weights for each subject matter. As noted in Table 3.7, weighting the two equally produces a composite score with $E\hat{\rho} = .817$; weighting

math twice as strongly as science produces a composite score with ($E\hat{\rho} = .775$), whereas weighting science twice as strongly as math produces a composite score with ($E\hat{\rho} = .816$).

UNBALANCED DESIGNS

The designs we have considered up to this point are balanced in the sense that the cells in the design have the same number of scores. The $p \times t \times r$ G study in which 20 persons submit four tasks (lesson plans) that are each scored by two raters is balanced. Similarly, a $p \times (t{:}r)$ design in which each person submits eight tasks (lesson plans) and four are scored by one rater and the other four are scored by a second rater may also be considered balanced. An "unbalanced" version of this second design would have an unequal number of levels, such as five lesson plans scored by one rater and three lesson plans scored by the second rater. Lack of balance may also arise when data are missing, such as a teacher failing to submit one of his or her lesson plans. Such missing data would create unbalancedness in both the $p \times t \times r$ and $p \times (t{:}r)$ designs just described.

Unbalancedness is essentially a form of missing data—whether planned (such as in the unbalanced nested design described) or unplanned (such as student absenteeism or failure to complete an activity). The presence of these missing data causes substantial problems in estimation of variance components. Only in balanced designs is variance component estimation as simple as setting expected mean squares equal to observed mean squares and solving for estimated variance components (the usual ANOVA approach). In unbalanced designs, the situation is more complicated. As pointed out by McCulloch and Searle (2001), there is no unique set of sums of squares as there is with balanced data and, consequently, "no unique set of [equations equating variance component estimates and functions of mean squares] and no unique estimators" (p. 173).

To address this issue, nearly 60 years ago, Henderson (1953) developed estimators based on ANOVA-like procedures, and these estimators are often used in specialty software available today (discussed in the Generalizability Study Programming Options section; see also extensive discussion in Brennan, 2001). Later, Rao (1971a, 1971b, 1972) developed a minimum norm quadratic unbiased estimation (MINQUE) strategy and its variants (e.g., MINQU(0), I-MINQUE), available in software packages such as SAS (SAS Institute Inc., 2010), to estimate variance components for unbalanced designs. More recently, likelihood-based methods (maximum likelihood estimation, restricted maximum likelihood estimation) and methods based on modern Bayesian simulation strategies such as Markov Chain Monte Carlo estimation (see, e.g., Gelman, Carlin, Stern, & Rubin, 2004) have become widely available in popular statistical packages such R (R Development Core Team, 2010), SAS (SAS Institute Inc., 2010), and SPSS (SPSS, Inc., 2010). Several resources are available for guidance on choosing an estimation procedure (Brennan, 2001; McCulloch & Searle, 2001; Searle, 1987; Searle, Casella, & McCulloch, 1992; Shavelson & Webb, 1981).[3]

VARIANCE COMPONENT ESTIMATION: SAMPLING VARIABILITY

As with any sample statistic, variance component estimates are subject to sampling variability. Early in the life of G theory, Cronbach et al. (1972, p. 49) raised the concern of instability of variance component estimates, especially with the modest sample sizes often used in G studies (see also Gao & Brennan, 2001, for empirical evidence of such instability).

Exact standard errors for variance component estimates are generally unavailable because of the inability to derive exact distributions for variance component estimates (see Searle, 1971). Satterthwaite (1941, 1946) and Ting, Burdick, Graybill, Jeyaratnam, and Lu (1990) developed procedures for obtaining

[3]Although ANOVA methods are still the most commonly used methods in G studies, the broader statistical community has largely moved away from them in favor of likelihood methods. Brennan (2001) promoted ANOVA-type estimators over likelihood methods largely because they (ANOVA methods) provide unbiased estimates, whereas likelihood methods require "suspect" assumptions. That said, Searle et al. (1992, p. 221) question whether it is appropriate to consider unbiasedness in random effects contexts. McCulloch and Searle (2001, pp. 173–174) make an even stronger case against ANOVA estimators, going so far as to summarize with "a consequence of all this is that ANOVA estimation of variance components is losing some (much) of its popularity."

approximate confidence intervals based on ANOVA methods (see also Burdick & Graybill, 1992), but these methods require the strong assumption of multivariate normality of score effects, which is usually untenable with a small number of conditions for each facet.

Two approaches to estimating sampling variability that invoke no distributional form are the jackknife (Tukey, 1958) and the bootstrap (Efron, 1982; Efron & Tibshirani, 1986). The jackknife involves taking n jackknife samples, each of size $n - 1$ and each omitting a different observation from the original data. The sample statistic of interest, then, is calculated for each of the jackknife samples, and the distribution of jackknife statistics is taken to represent sampling variability. The bootstrap is similar to the jackknife in that it involves drawing from the original sample a large number of bootstrap samples, each of which shares the same dimension as the original sample. The bootstrap differs from the jackknife in that each bootstrap sample consists of observations drawn with replacement from the original sample.

Extending these procedures to obtain estimated standard errors for variance components is not straightforward for multidimensional datasets, however (Brennan, 2001). For example, in a $p \times i$ (Persons $\times$ Items) design it is possible to obtain bootstrap samples by sampling persons with replacement but not items; by sampling items with replacement but not persons; by sampling persons and items with replacement; or by sampling persons, items, and residuals with replacement. Bootstrap estimates generated under these various strategies vary, often a great deal (Brennan, Harris, & Hanson, 1987; Tong & Brennan, 2007; Wiley, 2001). Wiley (2001) showed this inconsistency of results to be a function of two factors: (a) the resampling strategy used to generate them and (b) incongruence between the random effects specifications of a given G theory model and the treatment of effects as fixed in a given bootstrap strategy. As an example of the latter, consider a $p \times i$ random effects design. The creating of bootstrap samples by sampling persons with replacement (but treating items as fixed) violates the specification of item effects as random. Similarly, creating bootstrap samples by randomly sampling items with replacement (with no corresponding resampling of persons) treats persons as fixed. Wiley (2001) provided guidance on how to use the results from different bootstrap strategies to produce unbiased estimates of variance components and their sampling distributions for one-facet designs. Tong and Brennan (2007) provided further guidance for more complex designs.

GENERALIZABILITY STUDY PROGRAMMING OPTIONS

Several popular computer packages and programs provide estimates of variance components in G studies. These include SAS (SAS Institute Inc., 2010), SPSS (SPSS, Inc., 2010), and R (R Development Core Team, 2010). Several programs have been developed specifically for G theory, the most prominent of which are GENOVA (GENeralized analysis Of VAriance; Brennan, 2001; Crick & Brennan, 1983) and EduG (Cardinet, Johnson, & Pini, 2009). GENOVA handles complete balanced designs; its sister programs urGENOVA and mGENOVA provide applications specific to unbalanced designs (resulting from nesting and missing data) and multivariate designs, respectively. EduG includes a graphical user interface for easier use, although it cannot accommodate the range of models possible through GENOVA. That said, EduG does offer one modeling option unavailable through GENOVA—the program is flexible enough to specify different sources of variability as the measurement object (see Cardinet et al., 2009, for the advantages of this flexibility). Each application is available at no cost on its publisher's website.

OTHER ISSUES IN GENERALIZABILITY THEORY

We briefly mention additional issues (along with references) that are pertinent to G theory:

- Principle of symmetry: This principle holds that any one of the facets of a typical G study design may serve as the object of measurement (see Cardinet, Tourneur, & Allal, 1976, 1981).
- Generalizability of group means (important in cases in which groups of people—rather than individual people—serve as the objects of

measurement): This is one application of the principle of symmetry (see Kane & Brennan, 1977, for a detailed presentation of the use of G theory for group measurement; see also Webb, Shavelson, & Steedle, 2012).

- Nonconstant error variance for different true scores: In this chapter, variability of any given effect has (implicitly) been treated in this chapter as constant across the set of n_p true scores. This need not be the case. One might reasonably expect the population variability to vary systematically with individuals' true scores (see Brennan, 2001, for a discussion of the problems of, and solutions for, nonconstant error variability).
- Hidden (or implicit) facets: Sources of variance not explicitly modeled in a generalizability analysis are present in any study (Cronbach, Linn, Brennan, & Haertel, 1997; see also Shavelson, Ruiz-Primo, & Wiley, 1999).
- Linking G theory and item response theory: Providing a single model that incorporates these two distinct psychometric approaches has been of interest for many years (various strategies have been proposed by Bock, Brennan, & Muraki, 2002; Briggs & Wilson, 2007; Kolen & Harris, 1987; and Marcoulides & Drezner. 2000).

CONCLUSION

G theory offers a framework for researchers to specify a comprehensive sampling frame that considers a priori the potential sources of error in a measurement. The G study and D study allow the researcher to isolate the main sources of error variance, and in doing so they can be used to model how to maximize score reliability for a decision that must be made as well as how to choose an optimal design given resource constraints.

In this way, the contribution of G theory is much greater than simply the statistical mechanics for decomposing variance while estimating a reliability-like coefficient. It provides tremendous flexibility to accommodate designs with such issues as nested facets, unbalanced facets, and norm- versus criterion-referenced decisions. It allows for group reliability measurement as well as reliability measurement for multivariate composites. As such, it remains one of the major frameworks available to the applied psychometrician.

References

Bock, R. D., Brennan, R., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26,* 364–375.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press.

Brennan, R. (2001). *Generalizability theory: Statistics for social science and public policy*. New York, NY: Springer-Verlag.

Brennan, R., Harris, D., & Hanson, B. (1987). *The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts* (ACT Research Report Series 87-7). Iowa City, IA: American College Testing Program.

Briggs, D., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44,* 131–155. doi:10.1111/j.1745-3984.2007.00031.x

Burdick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components.* New York, NY: Dekker.

Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying generalizability theory using EduG*. New York, NY: Routledge.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*, 119–135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18,* 183–204.

Crick, J. E., & Brennan, R. (1983). *Genova: A generalized analysis of variance system* [Computer software and manual]. Iowa City: University of Iowa. Retrieved from http://www.education.uiowa.edu/casma

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373. doi:10.1177/0013164497057003001

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. *Annales de l'Institut Henri Poincare, 75*, 95–158.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611970319

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1,* 54–75.

Fyans, L. (1977). *A new multiple level approach to cross-cultural psychological research.* Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign.

Gao, X., & Brennan, R. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14,* 191–203. doi:10.1207/S15324818AME1402_5

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis.* Boca Raton, FL: Chapman & Hall.

Gleser, G. C., Green, B. L., & Winget, C. N. (1978). Quantifying interview data on psychic impairment of disaster survivors. *Journal of Nervous and Mental Disease, 166*, 209–216. doi:10.1097/00005053-197803000-00007

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics, 9,* 226–252.

Joe, G. W., & Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika, 41,* 205–207.

Kane, M. T., & Brennan, R. (1977). The generalizability of class means. *Review of Educational Research, 47*, 267–292.

Kolen, M. J., & Harris, D. J. (1987, April). *A multivariate test theory model based on item response theory and generalizability theory*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement, 54,* 3–7.

Marcoulides, G. A., & Drezner, Z. (2000). A procedure for detecting pattern clustering in measurement designs. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 287–300). Stamford, CT: Ablex.

McCulloch, C. A., & Searle, S. R. (2001). *Generalized, linear, and mixed models.* New York, NY: Wiley.

Rao, C. (1971a). Estimation of variance and covariance components—Minque theory. *Journal of Multivariate Analysis, 1,* 257–275. doi:10.1016/0047-259X(71)90001-7

Rao, C. (1971b). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis, 1,* 445–456. doi:10.1016/0047-259X(71)90019-4

Rao, C. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association, 67,* 112–115.

R Development Core Team. (2010). *R: A language and environment for statistical computing.* Vienna, Austria: Author.

SAS Institute, Inc. (2010). *The SAS system for Windows release 9.2.* Cary, NC: Author.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6*, 309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics, 2,* 110–114.

Schalock, H. D., Schalock, M. D., & Ayres, R. (2006). Scaling up research in teacher education: New demands on theory, measurement, and design. *Journal of Teacher Education, 57*, 102–119.

Searle, S. (1971). *Linear models.* New York, NY: Wiley.

Searle, S. (1987). *Linear models for unbalanced data.* New York, NY: Wiley.

Searle, S., Casella, G., & McCulloch, C. (1992). *Variance components.* New York, NY: Wiley.

Shavelson, R., Ruiz-Primo, M., & Wiley, E. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36,* 61–71.

Shavelson, R., & Webb, N. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133–166. doi:10.1111/j.2044-8317.1981.tb00625.x

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Snow, R. E., & Wiley, D. E. (Eds.). (1991). *Improving inquiry in social science: A volume in honor of Lee J. Cronbach.* Mahwah, NJ: Erlbaum.

SPSS, Inc. (2010). *SPSS base 19.0 for Windows user's guide.* Chicago, IL: SPSS, Inc.

Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S., & Lu, T. C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computation and Simulation, 35*, 135–143.

Tong, Y., & Brennan, R. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement, 67,* 804–817. doi:10.1177/0013164407301533

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics, 29*, 614.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 81–124). Amsterdam, the Netherlands: Elsevier.

Webb, N. M., Shavelson, R. J., & Steedle, J .T. (2012). Generalizability theory in assessment contexts. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 132–149). New York, NY: Routledge.

Wiley, E. (2001). *Bootstrap strategies for variance component estimation: Theoretical and empirical results.* Unpublished doctoral dissertation, Stanford University, Stanford, CA.

CHAPTER 4

TEST VALIDITY

Stephen G. Sireci and Tia Sukin

In educational and psychological testing, the term *validity* refers to the appropriateness and usefulness of a test for a particular purpose. In this sense, validity refers to an aspect of test quality; however, the concept extends beyond the test itself. As stated in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Thus, when evaluating a test, the test itself is not what is evaluated or validated; rather, the interpretations or decisions derived from test scores are what must be validated. Therefore, the concept of validity is comprehensive and refers not only to test characteristics but also to the appropriateness of test use and to the accuracy of the inferences made on the basis of test scores.

In this chapter, we provide a brief overview of validity theory, focusing on its evolution from the early 20th century to the current era. Next, we discuss test validation—the process of gathering and analyzing evidence to evaluate the appropriateness and utility of a test for a particular purpose. Our discussion of validation focuses on the five sources of validity evidence specified in the aforementioned *Standards* (AERA et al., 1999).[1] Specific research methods and statistical procedures that can be used to gather and analyze validity evidence are described. We conclude with a discussion of some of the major unresolved issues in validity theory and test validation.

Before proceeding, we must distinguish between *validity* and *validation*. As mentioned earlier, the former concept refers to the degree to which an assessment fulfills its intended purpose and test results are appropriately interpreted. *Validation,* however, refers to the process of gathering and reporting evidence to evaluate the use of a test for a particular purpose. In this chapter, we discuss both validity theory and test validation.

VALIDITY THEORY: PAST AND PRESENT

The history of formal standardized testing dates back at least to the civil service exams used in China around 165 B.C. (Teng, 1943). However, the modern era of standardized testing is typically traced to the work of Binet, who was commissioned to ensure that no child could be denied education in the Parisian school system without formal assessment (Binet, 1905; Binet & Henri, 1899). Binet's work was influential and soon led to intelligence testing and IQs (Terman & Childs, 1912; Terman et al., 1915). Standardized tests became popular because they were seen as an objective means for measuring unobservable psychological characteristics. The use of tests increased rapidly during World Wars I and II when hundreds of thousands of recruits needed to be evaluated and assigned to various positions.

[1]At the time of this writing, the *Standards* are undergoing revision. However, the current draft of the revision retains the categorization of validity evidence into the five sources we describe in this chapter.

DOI: 10.1037/14047-004
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

Use of tests on such a large scale generated research and debate on the utility and accuracy of test results and their interpretation. The concept of validity soon emerged to characterize concerns regarding test utility and the accuracy of test results.

Early Definitions of Validity

Given that psychological characteristics are generally unobservable, the earliest definition of validity, and one that still perseveres today, is the degree to which a test measures what it purports to measure (e.g., Garrett, 1937; Smith & Wright, 1928). This definition highlights the fact that in addition to defending the utility of a test, psychologists were also often defending the existence of the something a test was measuring. In educational and psychological testing, that something is called a *construct,* which as Cronbach and Meehl (1955) pointed out "is some postulated attribute of people, assumed to be reflected in test performance" (p. 283). Hence, validity theory and construct theory (providing evidence that the psychological attribute being measured actually exists) are closely intertwined. Although the notion of a construct was not formally linked with psychological testing until the mid-20th century (APA, 1954; Cronbach & Meehl, 1955), the notion of demonstrating that a test measures what it was designed to measure represented the first formulation of a validity theory.

First Validation Endeavors

The degree to which a test measured its theoretical something gave early psychometricians a theoretical framework for validating tests. How could test specialists at the beginning of the 20th century, however, provide evidence that tests were measuring what they purported to measure? As with current test specialists, they relied on the statistical procedures available at the time. In 1896, Karl Pearson published the formula for the correlation coefficient, which allowed quantification of the degree to which two variables relate to one another in a linear fashion. This revolutionary statistic was soon used to gauge the degree to which test scores correlated with other variables thought to measure the same construct. Typically, a nontest measure was considered to be the criterion of the measurement, and tests that provided scores that correlated highly with the criterion were considered valid. This perspective led Guilford (1946, p. 429) and others to claim "a test is valid for anything with which it correlates" (see also Bingham, 1937; Kelley, 1927; Thurstone, 1932). In keeping with this statistical validation of tests, correlations between test scores and criteria were put forward as validity coefficients. Ultimately, providing validity evidence that was based on relationships among test scores and criterion measures became known as *criterion-related validity evidence* (APA, 1966).

Pearson (1901) extended his work on correlation into principal component analysis, which is a statistical method for creating a subset of variables that represent the variation among a much larger set of variables. Spearman (1904) expanded on Pearson's components by hypothesizing underlying psychological traits that explained examinees' performance on tests. His methods of factor analysis were subsequently used to identify the latent traits that explained test performance. The techniques of factor analysis (which we explain in the section Validity Evidence Based on Internal Structure later in this chapter) were used by early 20th-century psychometricians, and are still used today, to provide evidence that a test is measuring the construct it is designed to measure. Guilford (1946) referred to such evidence as *factorial validity*.

Concerns Over Criterion and Factorial Validity

Although factor analysis and test–criterion relationships provided useful validity information, over time it became clear that such information was limited and often insufficient to support the use of a test for a particular purpose. Concerns over the limitations of a purely statistical approach to validation led many early test specialists to conclude that a more comprehensive strategy was required—one that focused on the degree to which a test fulfilled its purpose (e.g., Jenkins, 1946; Kelley, 1927; Pressey, 1920; Thorndike, 1931) and, in some cases, on the degree to which the content of a test adequately represents what is intended to be measured (Ebel, 1956, 1961; Lennon, 1956; Rulon, 1946).

The issue regarding the degree to which a test adequately represented what it was supposed to

measure led to the development of a new source of evidence for validating tests on the basis of the quality and appropriateness of the test content—a concept that was ultimately termed *content validity* (APA Committee on Test Standards, 1952; Cureton, 1951). *Content validity* refers to the degree to which the content of a test is representative of the targeted construct and supports the use of a test for its intended purposes (Ebel, 1956; Lennon, 1956; Sireci, 1998a). This type of evidence was thought to be particularly important for educational tests such as those used in achievement testing or in a credentialing context. Content validity evidence is usually gathered by having experts review test items and make judgments regarding the relevance of each item to the construct measured and the degree to which the items adequately and fully represent the construct (Sireci, 1998b).

Thus, by the middle of the 20th century consensus was growing that validation was a comprehensive process focused on specific uses and interpretations of test scores and multiple forms of evidence to support such uses. The awareness of the limitations of purely statistical perspectives on validation, a growing awareness of the importance of content validation, and a concern over the use of tests for purposes beyond their means led to a movement to develop consensus standards for developing and validating tests.

Validity and the Standards for Educational and Psychological Testing

By the mid-20th century, several different ideas about validity and different types of validity evidence existed. To foster consensus on these issues and practices, APA convened a commission that developed *Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal* (APA Committee on Text Standards, 1952). Shortly thereafter, AERA and NCME joined in, resulting in a joint effort that has produced standards for educational and psychological testing since 1954. The first version of these joint standards specified four types or attributes of validity: predictive validity, concurrent validity, content validity, and construct validity (APA, 1954). Subsequent versions of the standards were published in 1966, 1974, 1985, and 1999 (APA, 1966, 1974; AERA, APA, & NCME, 1985, 1999). Chapter 13 in this volume provides additional information about these test standards and their history.

The distinction between predictive and concurrent validity regards whether the test is designed to predict future performance or current performance. Predictive validity is used to describe the degree to which test scores are useful for predicting success on the job or in an educational setting such as college or graduate school. Concurrent validity describes how well test scores relate to an individual's standing on a current criterion. A common application of concurrent validity is the degree to which scores on a short form of a test are strongly associated with scores on a longer form of the test.

In Table 4.1, we provide a brief description of the different ways in which validity evidence was categorized in each version of these standards (Sireci, 2009). As is evident from Table 4.1, predictive validity and concurrent validity were subsumed under the more general term *criterion-related validity,* which as we described earlier refers to the relationships among test scores and external criteria. *Construct validity*, which was introduced in APA's 1954 *Technical Recommendations for Psychological Tests and Diagnostic Techniques* and expanded on by Cronbach and Meehl (1955), refers to the degree to which test scores represent an individual's standing on the theoretical construct the test is designed to measure. Because many test specialists consider construct validity to be the most general category of evidence (e.g., Messick, 1989), we elaborate on the concept in the next section.

Current Conceptualizations of Validity

Two concepts are central to understanding contemporary validity theory and test validation. The first is the unitary conceptualization of validity centered on construct validation; the second is the argument-based approach to validation.

Unitary conceptualization of validity. In describing construct validity, Cronbach and Meehl (1955) stated, "Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to

TABLE 4.1

Categorization of Validity Evidence Over Time in the *Standards*

Publication	Validity classifications
Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal (APA Committee on Test Standards, 1952)	*Categories:* predictive, status, content, congruent
Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, 1954)	*Types:* construct, concurrent, predictive, content
Standards for Educational and Psychological Tests and Manuals (APA, 1966)	*Types:* criterion related, construct related, content related
Standards for Educational and Psychological Tests (APA, 1974)	*Aspects:* criterion related, construct related, content related
Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985)	*Categories:* criterion related, construct related, content related
Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999)	*Sources of evidence:* content, response processes, internal structure, relations to other variables, consequences of testing

Note. APA = American Psychological Association; AERA = American Educational Research Association; NCME = National Council on Measurement in Education. From *The Concept of Validity: Revisions, New Directions, and Applications* (p. 26), by R. Lissitz (Ed.), 2009, Charlotte, NC: Information Age. Copyright 2009 by Information Age Publishing Inc. Adapted with permission.

be measured" (p. 282). Many test specialists argued that no criterion or content universe is entirely adequate, and so all validity was essentially construct validity (Loevinger, 1957; Messick, 1989). The logic underlying this argument is that all interpretations of test scores imply an underlying construct. Thus, over time, construct validity became more comprehensive. Messick (1989), for example, wrote, "Construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 17). Obviously, this definition implies that construct validity subsumes all other types of evidence.

Although the unitary conceptualization of validity stating that all validity is construct validity is intellectually compelling, not all test specialists have agreed with this perspective (e.g., Ebel, 1956, 1961), and some have argued that avoiding terms such as *content validity* will have a negative impact on validation practices (e.g., Sireci, 1998a; Yalow & Popham, 1983). Nevertheless, construct validity theory provides a helpful framework for evaluating the use of a test for a particular purpose. For example, Messick (1989) claimed, "Tests are imperfect measures of constructs because they either leave out something that should be included according to the construct theory or else include something that should be left out, or both" (p. 34). He suggested that test validation endeavors focus on (a) identifying sources of construct-irrelevant variance and (b) determining whether the construct is underrepresented. Most threats to the validity of test scores can be classified into one of these two general areas.

Argument-based approach. Kane (1992, 2006) borrowed from Cronbach's (1971, 1988) perspective of validation as evaluation to propose an argument-based approach to validation. In this approach, test users develop an interpretive argument, which specifies (a) how test scores will be used or interpreted and (b) the logical chain of inferences to support the interpretations or uses. On the basis of this chain of inferences, validity hypotheses are proposed and tested to establish a validity argument. The argument-based approach is similar to defending the use of a test in a courtroom. The idea is to present a preponderance of evidence that would support the use of a test for a particular purpose. This body of evidence should include evidence that the test is fulfilling its intended objectives and is not producing undesired consequences. The approach

does not specify a particular kind of validity evidence, and so it assimilates the traditional forms of content- and criterion-related validity and is consistent with a comprehensive construct validity perspective.

The current *Standards for Educational and Psychological Testing* (AERA et al., 1999) implicitly support the argument-based approach to validation. For example,

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . . . Ultimately, the validity of an intended interpretation . . . relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees. (AERA et al., 1999, p. 17)

To summarize our historical review of validity theory, the earliest validity theorists defined validity in terms of test and criterion relationships. Such relationships were used to demonstrate that tests measured what they were supposed to measure. As the consequences associated with tests grew, critics argued that more was needed to justify the use of a test for a particular purpose. These criticisms led to more comprehensive views of validity that require (a) evidence that the test is consistent with the theory on which it is based, (b) that it provides the desired types of information consistent with its intended purposes, and (c) that it does not result in negative, unintended consequences for individuals, groups, or society. Today, these concerns are addressed by the *Standards,* which support an argument-based approach to validating tests, centered on five sources of validity evidence.

Five sources of validity evidence. The *Standards* stipulate five sources of validity evidence "that might be used in evaluating a proposed interpretation of test scores for particular purposes" (AERA et al., 1999, p. 11). The sources are validity evidence based on (a) test content, (b) relations to other variables, (c) internal structure, (d) response processes, and (e) consequences of testing.

Given a particular use of a test, one source of evidence is likely to be more important than another. For example, validity evidence based on test content is likely to be particularly important in educational achievement testing. However, one source of evidence is not likely to be sufficient for a compelling validity argument, and in most cases, multiple sources of evidence are needed.

VALIDATION

The previous review of validity theory demonstrated that what needs to be validated is the use of a test for a particular purpose. Thus, validation starts with explication of the purpose of a test and ends with evidence bearing on the appropriateness of the interpretations based on test scores with respect to the specific purpose. Such evidence should be comprehensive and compelling so that the interpretations made on the basis of test scores can be supported. In this section, we focus on how such evidence can be gathered, analyzed, and evaluated using the *Standards*' five sources of validity evidence (AERA et al., 1999).

Validity Evidence Based on Test Content

The first step in developing a test is defining what is being measured. From a construct perspective, this is called *defining the construct;* from a content perspective, it is called *domain definition.* Regardless of nomenclature, defining the construct tested is also the first step in providing validity evidence based on test content because the definition of the construct has a direct impact on score interpretation and use.

Validity evidence based on test content can be categorized into four general areas: (a) construct definition, (b) construct relevance, (c) construct representation, and (d) appropriateness of test construction procedures (Sireci, 1998a). *Construct definition* refers to the appropriateness of how the domain to be tested is described and specified. In evaluating the construct definition, one evaluates

the test developer's general statements regarding what is being tested as well as the specifications that describe test content. In educational testing, test specifications are often in the form of a 2 × 2 table that indicates the content areas and cognitive skills measured, although other structures are also possible. Evaluators typically focus on whether any important areas are omitted from the specifications or whether superfluous areas are included.

In licensure, certification, and other employment contexts, the construct domain tested is often defined on the basis of a practice analysis or job analysis (Raymond, 2001). (See Volume 3, Chapter 19, this handbook for additional information on job and practice analysis.) Practice analyses are used in licensure and certification settings, whereas job analyses are used in noncredentialing contexts. These analyses survey or observe practitioners to define the most frequent and important tasks conducted on the job. On the basis of these frequency and criticality data, test specifications are derived to represent the job domain. These specifications represent the operational definition of the tested domain.

After evaluating the construct definition, a second important area is evaluating construct relevance. Here, validation focuses on whether each item on the assessment is relevant to the construct tested. A related area is construct representation, which investigates the degree to which the items and tasks on an assessment fully represent the intended construct and do not contain material irrelevant to the construct measured. Traditional content validity studies use subject matter experts to evaluate test items for relevance and representativeness (Crocker, Miller, & Franks, 1989; Martone & Sireci, 2009). Items that are not judged relevant to the domain are eliminated, and new items are added if the experts conclude that aspects of the domain are underrepresented. In evaluating construct representation, the relative proportion of items measuring different aspects of the domain is also appraised.

An evaluation of the appropriateness of the construct definition focuses on how well that definition captures the consensus understanding of the construct. An analysis of construct relevance focuses on the degree to which each item appropriately measures some aspect of the construct. An analysis of construct domain representation focuses on the degree to which the test as a whole adequately represents the domain. Finally, evaluation of the appropriateness of test construction procedures involves looking at the various item development, selection, scoring, and quality control procedures involved in constructing the assessment. Elements looked for in evaluating these procedures include training of item writers, qualitative and statistical criteria for selecting items, adequacy of scoring rubrics, screening out potentially biased items (sensitivity review; see Ramsey, 1993; Sireci & Mullane, 1994), and quality control checks to ensure accurate scoring.

In educational testing, a relatively new aspect of validity evidence based on test content is alignment. Alignment methodology arose from concerns about how well statewide educational tests aligned with statewide curricula. According to Webb (1997), *alignment* refers to "the degree to which expectations [i.e., standards] and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (p. 4). Much of the data gathered through an alignment study are similar to the types of data gathered through traditional content validity studies. However, many alignment studies are more complex and often involve first rating the benchmarks (educational objectives) within a curriculum framework before rating the items. Some alignment methods also evaluate the degree to which instruction is consistent with what is measured on a statewide assessment (Porter, Smithson, Blank, & Zeidner, 2007).

To summarize validity evidence based on test content, content-related evidence is typically gathered using subject matter experts who review test items and rate them with respect to their relevance and appropriateness for measuring the construct and with respect to the adequacy with which test content is congruent with the purpose of testing. Such analyses are important because they represent independent appraisals of what is intended to be measured.

Validity Evidence Based on Relations to Other Variables

Although criterion-related validity evidence has been established as insufficient to fully support claims of

validity for score interpretations and use, depending on the purpose of the test such evidence may be quite useful in building the overall validation argument. Validity evidence based on relationships between test scores and other variables extends beyond single test–criterion relationships and includes the analysis of the relationships of test scores with constructs that are expected to be positively related, negatively related, or unrelated. Confirmation of such theoretical relationships can reinforce the interpretations and uses that are intended to result from a score on a given instrument.

Two questions are helpful in guiding the collection of criterion-related validity evidence: (a) Is the rationale for selecting criterion variables and demonstrating their suitability appropriate and (b) are the patterns observed between the test scores and external variables consistent with prior expectations? Important issues to consider in gathering validity evidence based on relationships with external variables include (a) the type of information gained, (b) points to consider when selecting criterion variables, and (c) the potential threats to validation that may be addressed.

Selection of validation criteria. The selection of validation criteria is one of the most important tasks when constructing a validation argument that includes criterion-related evidence. Important considerations include the relevance, practicality, and reliability of each criterion. To the extent that the criteria are not relevant to the construct of interest, the validity argument will be weakened. Thus, the selection of the criteria on the basis of the hypothesized relationships stemming from the theory underlying the construct measured is imperative. Cronbach (1988) and Kane (2006) distinguished between the strong form of construct validation, in which selection of validation criteria is based on construct theory, and the weak form of construct validation, in which "any correlation of the test score with another variable is welcome" (Cronbach, 1988, p. 13).

Selecting criteria that are consistent with the theory underlying the construct is easier said than done. Valid external criteria are hard to obtain, expensive, or both. In other cases, valid external criteria may simply not exist. Often, collecting criterion data is impractical, and even when such data are gathered, they may be of questionable reliability or corrupted by biases (e.g., supervisors' ratings used in an employment setting or teachers' ratings in an educational setting).

In addition, criterion variables are often subject to statistical artifacts such as sampling error, weak to moderate reliabilities, and range restriction. Highly select samples underestimate validity by restricting the range of the observed scores (i.e., the standard deviation is smaller in the selected sample than it would be in the entire pool of examinees). Sackett, Borneman, and Connelly (2008) gave the example of a correlation of .50 between the criterion and a predictor in an unrestricted sample that drops to .33 when only the top 50% of the candidates are selected and thus included in the criterion measures. Moreover, an unreliable measurement of the criterion also results in an underestimation of the validity.

Basing judgments of validity on a poor data collection design is misleading because an unreliable criterion measure will diminish the correlation of the measure with the criterion. Fortunately, statistical corrections are available for these statistical artifacts (Sackett et al., 2008; Sackett & Yang, 2000).

Statistical artifacts aside, the development of a validity argument through the use of relationships with other variables can help to alleviate common threats. The degree to which criterion-related validity evidence supports a validity argument depends in large part on the reasonableness of the validation criteria. Next, we discuss the statistical methods used in analyzing criterion-related validity data. However, no matter how sophisticated the statistical analysis, if the results are to provide compelling evidence regarding validity, the criteria included in the analysis must be justified on both theoretical and technical grounds.

Correlation. The simplest index of test–criterion relationship is Pearson's product–moment correlation coefficient (r_{xy}), which is calculated by dividing the covariance between two variables of interest (e.g., a predictor variable and a criterion variable) by the product of their standard deviations. Correlation coefficients range from -1 to 1, where 1 indicates a perfect positive linear relationship and -1 indicates

a perfect negative linear relationship. As the correlation magnitude approaches zero, so too does the strength of this linear dependence.

Multitrait–multimethod correlations. Correlations can be used to assess both convergent and discriminant relationships. In fact, correlations are the focus of analysis in Campbell and Fiske's (1959) multitrait–multimethod approach to validation. Convergent relationships indicate that different methods of measurement of the same construct are positively and highly correlated with one another, whereas discriminant relationships indicate that dissimilar constructs are not highly correlated.

Campbell and Fiske (1959) suggested that one comprehensive approach to evaluating validity would be to include different constructs (referred to as *traits*) in the analysis as well as qualitatively different measures of the same construct. For validity to be supported, one would expect high correlations for the same construct on different measures (convergent validity) and noticeably lower correlations among measures of different constructs. Their framework specified different measures of the same construct as different methods of measuring a trait, hence the description *multitrait–multimethod.*

To evaluate convergent and discriminant validity and to investigate the presence of construct-irrelevant method variance, Campbell and Fiske (1959) proposed arranging correlations among multiple measures of multiple constructs into a multitrait–multimethod matrix. This matrix is an arrangement of correlations such that the correlations among the different traits are stratified by the different methods. Convergent validity is evaluated by inspecting the monotrait–heteromethod correlations (same trait measured by different methods), and discriminant validity is evaluated by inspecting the heterotrait–monomethod (different traits measured by the same method) and heterotrait–heteromethod correlations. A sound validity argument based on this approach would exhibit large and statistically significant monotrait–monomethod correlations that were substantially larger than the heterotrait correlations. As Campbell and Fiske pointed out, "Tests can be invalidated by too high correlations with other tests from which they were intended to differ" (p. 81).

Data considerations in correlation analysis. Whenever one is interpreting validity evidence on the basis of correlations, the nature of the data should be considered. The use of correlation assumes a linear relationship between the predictor and the criterion. If a nonlinear relationship is present, a Pearson correlation coefficient will underestimate the relationship and result in misleading conclusions.

Another notable problem in correlation analysis is restriction in range, which occurs when the sample used to calculate a correlation is more homogeneous (contains less variability) than the population to which the validity inference is to generalize. Restriction in range is a common occurrence in predictive validity studies, such as when grades in college are used to evaluate the predictive power of college admissions tests. Because only students who were admitted to college have data on the criterion (college grades), they represent a more restricted sample than the entire population of college applicants who took the test (i.e., many of these examinees were not accepted into college). This restricted variability attenuates (weakens) the correlation.

Disattenuating correlations for restriction in range. Fortunately, when there is information regarding the variability of the predictor in the population, correlations can be disattenuated for restriction in range. Given the observed standard deviations on the predictor ($\sigma_{\tilde{x}}$) and criterion ($\sigma_{\tilde{y}}$) for the restricted samples, along with the standard deviations for the population on the predictor (σ_x) and the correlation between the predictor and the criterion ($\rho_{\tilde{x}\tilde{y}}$), the standard deviation of the criterion for the entire pool of applicants (σ_y) can be estimated as

$$\sigma_y = \sigma_{\tilde{y}} \sqrt{\frac{\sigma_x^2 \rho_{\tilde{x}\tilde{y}}^2}{\sigma_x^2} + 1 - \rho_{\tilde{x}\tilde{y}}^2}. \qquad (4.1)$$

The standard deviation estimated in Equation 4.1 can then be used to estimate the disattenuated correlation (ρ_{xy}) between the predictor and criterion variables by

$$\rho_{xy} = \frac{\sigma_x \rho_{\tilde{x}\tilde{y}}}{\sqrt{\sigma_x^2 \rho_{\tilde{x}\tilde{y}}^2 + \sigma_{\tilde{x}}^2 - \sigma_{\tilde{x}}^2 \rho_{\tilde{x}\tilde{y}}^2}}. \qquad (4.2)$$

Disattenuating correlations for unreliability. Another disattenuation formula can be applied to

account for the imperfect reliability to which the criterion is measured, if information on the reliability of the criterion variable was available. This formula is

$$\rho_{\tilde{x}y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy}}}, \tag{4.3}$$

where $\rho_{\tilde{x}y}$ is the disattenuated correlation, ρ_{xy} indicates the uncorrected correlation, and ρ_{yy} represents the reliability of the criterion measure.

Correlations can be corrected for statistical artifacts such as restriction in range and unreliability of the criterion measure. In most cases, reporting both corrected and uncorrected correlations is wise (AERA et al., 1999, pp. 21–22). In addition, structural equation modeling (see Chapter 5, this volume) can be used to automatically adjust for unreliability of observed variables used to measure latent variables.

Multiple regression. Multiple regression is also used to gather validity evidence based on the relations of test scores with other variables. In addition to indexing the strength of a test–criterion relationship, multiple regression allows for gauging the predictive accuracy of test scores as well as the relative predictive utility of test scores when used in conjunction with other predictor variables.

Regression analysis fits a line that minimizes the squared deviation of the observed values on the criterion (y) from the values predicted from the regression line ($\hat{y}$):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \tag{4.4}$$

where y_i represents the score on the criterion for individual i, β_0 indicates the intercept of the regression line, β_p represents the weight or slope (also called the regression coefficient) associated with the first of p predictors, x_{1i} represents the score of person i on the first predictor, and ε_i represents the residual for individual i. Each regression coefficient represents the expected increase in the criterion associated with a one-unit increase in the predictor, holding the other predictors constant.

The regression analogue to the correlation coefficient is the multiple correlation coefficient (R), which reflects the correlation between the derived linear combination of the predictors and the criterion ($r_{y\hat{y}}$). The square of the multiple correlation coefficient is denoted R^2 and represents the proportion of variance in the criterion variable accounted for or explained by the linear combination of the predictors.

In addition to R and R^2, each element in a regression equation can provide validity evidence. The statistical significance of a predictor (e.g., test score) can be evaluated by testing the regression coefficient for statistical significance.

In many cases, however, a validity investigation may seek to understand whether test scores add to predictive accuracy above and beyond other predictors that may be available. Several indices are available to evaluate the utility of a single predictor in a multiple regression analysis including squared semipartial correlations, squared partial correlations, and the relative Pratt index (Thomas, Hughes, & Zumbo, 1998).

The squared semipartial correlation represents the proportion of total variance in the criterion variable accounted for by the predictor variable above and beyond that accounted for by the other predictors. A straightforward way to compute the squared semipartial correlation is to run a regression analysis twice, with and without the predictor variable of interest. For example, suppose a regression analysis involved two predictors: x and z. If $R^2_{y \cdot xz}$ represents the squared multiple correlation when both x and z are regressed on y and $R^2_{y \cdot z}$ represents the squared correlation when only z is regressed on y, the squared semipartial correlation for predictor x ($r^2_{y(x \cdot z)}$) is

$$r^2_{y(x \cdot z)} = R^2_{y \cdot xz} - R^2_{y \cdot z}. \tag{4.5}$$

Instead of focusing on the total variance in the criterion, the squared partial correlation focuses on the criterion variance unexplained by the other predictors. It refers to that portion of the unexplained variance that the predictor of interest does account for. If $R^2_{y \cdot xz}$ represents the squared multiple correlation when both x and z are regressed on y and $R^2_{y \cdot z}$ represents the squared correlation when z is regressed on y, the squared semipartial correlation for x ($r^2_{xy \cdot z}$) is

$$r^2_{xy \cdot z} = \frac{R^2_{y \cdot xz} - R^2_{y \cdot z}}{1 - R^2_{y \cdot z}}. \tag{4.6}$$

Because the equation for the semipartial correlation is the numerator in the equation for the partial correlation, the partial correlation (and its square) will always be larger than the semipartial correlation (and its square), unless the predictor is completely uncorrelated with the other predictors (in which case the semipartial and partial correlations would be equal). In either case, in interpreting these correlations one must remember that they are not indicators of the strength of the predictor, but rather an indication of the proportion of criterion variance that is explained after considering the contributions of all other variables in the equation.

The relative Pratt index is sometimes used instead of squared semipartial or partial correlations for determining the relative importance of variables because they are sometimes simpler to interpret because the sum of all Pratt indices (d_x) is 1 when all predictors are included. The equation for the relative Pratt index is

$$d_x = \frac{\hat{b}_x r_x}{R^2}, \tag{4.7}$$

where $\hat{b}_x$ represents the estimated regression beta weight, r_x represents the Pearson correlation between the predictor and criterion variables, and R^2 represents the total variance accounted for by the complete model.

Differential predictive validity. Regression and multiple regression can also be used to evaluate potential test bias. Specifically, if test scores are used to predict future performance, such as the use of college admissions tests to predict academic success in college, the degree to which the prediction is consistent and accurate across subgroups of examinees is relevant to the question of test bias. If separate regression lines are computed for subgroups of examinees, the regression weights and intercepts can be tested for statistically significant differences that may indicate bias (Wainer & Sireci, 2005). However, in many cases there are too few members of one or more subgroups for stable estimation of regression coefficients. An alternate strategy is to fit a single regression line for all groups and evaluate the errors of prediction to see whether they are consistent across subgroups of examinees (e.g., minority vs. nonminority examinees). If errors systematically differ (result in patterns of over- and underprediction), the use of the regression equation for all groups of interest may be questioned (Sireci & Talento-Miller, 2006). However, we should caution that in using these approaches, if group means differ, the results may be misleading.

Data considerations in multiple regression. Range restriction and unreliability of measures are also issues of concern in multiple regression, although disattenuation formulas such as those described earlier can be used to ameliorate their effects. As with correlation analysis, regression also assumes identically and independently distributed errors. However, an additional concern in regression analysis is multicollinearity, which refers to the case when two or more predictor variables are highly correlated. If the predictors are highly correlated, the regression coefficients may be unstable.

Hierarchical linear modeling. Although multiple regression has been widely used in validity analyses, researchers have pointed out that in many cases, validity analyses involve a nested structure that violates the independence-of-observations assumption in regression. For example, individuals are often nested within classrooms, schools, districts, teams, or some other functional unit. An alternative, multilevel modeling, has been proposed to address this problem, the most popular form being hierarchical linear modeling (Bryk & Raudenbush, 1992).

Hierarchical linear modeling involves calculating different predictor equations for the multiple groups or levels, thus treating the levels as fixed effects in the model. The first level consists of the model for the individual, and the subsequent levels pertain to the models associated with the grouping-level variables, such as schools, school districts, states, or other levels within which individuals can be nested. Korbin (2010), for example, applied hierarchical linear modeling to demonstrate the utility of a multilevel model to understand the relationship of institutional characteristics (e.g., size, type) in relationship to the validity of the SAT (among other variables) for predicting 1st-year college grade point average. When assessing the SAT for its predictive validity for different kinds of institutions, 1st-year college grade point averages become nested within

the institutional variables. Thus, multiple levels are required to model the impact of SAT scores in predicting 1st-year college grade point average. Generally, in the second and subsequent levels of a hierarchical linear model, the intercepts and slopes are what become the dependent variables in the next level of the model. The generation of the regression equations in this manner incorporates the variation across the individual subjects within a level, which cannot be accomplished by conducting separate regression equations.

It is important to note that hierarchical linear modeling requires larger data demands than a typical multiple regression, which limits its applicability. For example, Hox (1995) suggested a minimum of 20 groups for the highest level and 100 groups when variance components are to be estimated with low standard errors.

Groups as criteria: Experimental and quasi-experimental designs. In many validation endeavors, instead of evaluating prediction, the goal is to rule out rival hypotheses in building the validity argument. In such cases, experimental or quasi-experimental studies may be used. The difference between experimental and quasi-experimental designs is whether participants are randomly assigned to conditions in the experiment. Thus, when sex or ethnicity is the grouping variable, a true experimental design is impossible because these variables cannot be randomly assigned. Experimental designs may also use a repeated-measures format if the order of treatment was counterbalanced.

Experimental and quasi-experimental designs are typically used to test specific validity hypotheses. In these cases, test scores are the dependent variable on which the groups are compared. Validity hypotheses may specify statistically significant differences across two or more groups or may hypothesize no difference. For example, if a test is designed to measure sensitivity to instruction, students can be randomly assigned to instruction and noninstruction groups and then given the test after the students in the experimental group completed the instruction. The validity hypothesis would be that students in the experimental group would perform better on the test. A validity hypothesis predicting no difference across groups would be exemplified by testing programs that administer parallel forms of a test delivered either over a computer or via traditional paper-based testing. Some studies of computer-based and paper-based test comparability have used experimental designs (Bennett et al., 2008; Kim & Huynh, 2007), and others have used quasi-experimental designs because of difficulties in making random assignments (Glasnapp, Poggio, Carvajal-Espinoza, & Poggio, 2009; Puhan, Boughton, & Kim, 2007). Quasi-experimental designs in this context typically involve matching students who took a test on computer or paper on demographic and academic variables (e.g., prior year's test scores). Quasi-experimental designs have also been used to evaluate the effects of test accommodations on students' performance on educational tests (Sireci, Scarpati, & Li, 2005).

Validity generalization and meta-analysis. Before concluding our discussion of criterion-related validity evidence, we should mention two other topics that cut across statistical procedures—validity generalization and meta-analysis. *Validity generalization* is a process through which validity evidence from different studies is quantitatively merged to determine the degree to which statistical evidence of validity gathered in one situation extends to different settings, such as organizations, time periods, jobs, and geographical areas. Quantitative methods for validity generalization include meta-analyses and Bayesian techniques. Typically, the goal is to use the results of existing criterion-related studies and apply them to new situations, new settings, and new populations of examinees.

Meta-analysis refers to statistical summary of several studies on a particular topic. With respect to validity generalization, meta-analysis is used to summarize the results of several criterion-related validity studies so that the general trend and magnitude of prediction or utility can be ascertained. The use of meta-analysis for supporting validity generalization involves the collection of data that span situations, settings, and populations. The more areas to which the results are shown to generalize, the greater the confidence in applying the criterion relationships to unique situations. Here, it is especially important to remember that each observed validity statistic is a

sample statistic and thus subject to artifactual variance because of differential restriction in range, small sample sizes, and differential criterion unreliability (Schmidt, 1988).

Meta-analysis can involve several approaches to summarizing validity data. For example, when a test score is used as a single predictor for a specific criterion, the median test–criterion correlation or, more commonly, a weighted average correlation, in which the weighting is based on the sample size, can be reported. Effect size measures may also serve as the summary index. When the criterion is a grouping variable, Cohen's (1988) delta may be used to summarize the results across studies (e.g., weighted average delta). In correlation analysis, the weighted average squared correlation may serve as a summary index. Meta-analysis also involves drawing confidence intervals or other information regarding expected variability so that the degree to which the results might replicate can be provided.

The *Standards* (AERA et al., 1999) point out that the different contexts of the studies involved in validity generalization should be considered. They recommend that the characteristics of all studies summarized be reported and any substantive differences across studies be considered. For example, if subsets of studies in a meta-analysis differ according to some feature (e.g., job family, region of the country, or other moderator variable), the *Standards* recommend reporting separate effect-size estimates for each feature (AERA et al., 1999, p. 22).

Validity Evidence Based on Internal Structure

The term *internal structure* refers to the dimensionality or underlying factor structure of an assessment. The theory used to develop a test will often hypothesize a specific dimensionality. For example, an assessment of self-concept may hypothesize separate dimensions for academic self-concept and social self-concept. Statistical analysis of test data can determine how many dimensions (or factors) are needed to characterize the variation in the data. The degree to which these empirically derived dimensions are congruent with the theorized dimensionality of the construct is one way to evaluate how well the test morphologically represents the construct.

Validity evidence based on internal structure can come from many different sources, including analysis of (a) internal consistency, (b) dimensionality, and (c) measurement invariance. In some cases, investigations of internal structure simply seek to justify the use of a particular scoring model, such as when unidimensional item response theory (IRT) models are used to calibrate items and provide scores for examinees. In other cases, the hypothesized multidimensionality of a construct is empirically tested through factor analysis or other means.

Important issues to consider and clarify in gathering validity evidence based on the internal structure of an assessment include (a) the type of information gained from collecting this kind of evidence, (b) the scoring model for the assessment, (c) the declaration of dimensionality, and (d) the decision to report subtest scores, composite scores, or both. Furthermore, validity investigations based on internal structure can be analyzed across subgroups of examinees to evaluate whether the test as a whole, subtests, or individual items are invariant across relevant subgroups of test takers.

Understanding dimensionality. Some assessments are intended to be unidimensional, and others are designed to be multidimensional. A *dimension* is a homogeneous continuum that accounts for variation in examinees' responses to test items. Analysis of internal structure involves some type of comparison of the hypothesized and observed dimensionalities. The type of scores derived from the assessment, such as a composite score, subtest scores, or score profile, must also be considered. For example, the *Standards* state,

> It might be claimed . . . that a test is essentially unidimensional. Such a claim could be supported by a multivariate statistical analysis, such as a factor analysis, showing that the score variability attributable to one major dimension was much greater than the variability attributable to any other identified dimension.. . . When a test provides more than one score, the distinctiveness of the separate scores should be demonstrated, and the interrelationships of those scores should

> be shown to be consistent with the construct(s) being assessed. (AERA et al., 1999, p. 20)

Thus, validity evidence based on internal structure should be reported to defend the types of scores provided by the test as well as the theoretical interpretation of those scores.

Statistical methods for evaluating internal structure. Numerous methods exist for evaluating the dimensionality of an assessment. These methods range from reporting estimates of internal consistency reliability to more sophisticated methods that are based on factor analysis, multidimensional scaling, and structural equation modeling. Next, we provide descriptions of some of the most relevant approaches.

Dimensionality analyses. Supporting claims of dimensionality for an assessment relative to the construct or content domain of interest is important. The following sections describe several methods for conducting dimensionality analyses, including exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling (MDS), and IRT residual analysis. Readers with interest in factor analysis and IRT should consult Chapters 5 and 6 in this volume.

Exploratory factor analysis. Pearson (1901) developed principal-components analysis to reduce large sets of correlations (or covariances) to a smaller number of components that represented the majority of the variation observed in the full set of correlations. The method derives a first component by finding a weighted linear combination of variables that account for the most observed variance among the correlations. A second component is then derived that accounts for the most remaining residual variance. This process is repeated until most of the variation is accounted for and any remaining components account for trivial variance.

Although principal component analysis is sometimes used to evaluate the internal structure of an assessment, EFA is more common and more appropriate. Spearman (1904) expanded principal component analysis into factor analysis by partitioning the total observed variance into common variance and unique variance. The components derived using factor analysis are called *factors,* and rather than representing observed variance, they represent the shared variance among the items. In the context of the analysis of test structure, each factor in an EFA refers to a hypothesized latent variable that explains examinees' responses to test items. (A full description of principal component analysis and EFA is beyond the scope of this chapter, but readers are referred to Chapter 5, this volume.) Essentially, the EFA model represents the response of an individual i to an item j as

$$m_{ij} = \sum_k a_{jk} f_{ik} + e_i, \tag{4.8}$$

where k is a particular factor. Thus, a person's response to an item is determined by the item's loading on a factor (a_{jk}) and a person's factor score (f_{ik}). The matrix algebra version of the EFA formula is more illuminating regarding how it can be used to evaluate internal structure because it describes the analysis in terms of the matrix of correlations among the variables:

$$R = VLV, \tag{4.9}$$

where R = the matrix of correlations among the variables (e.g., test items), V = a matrix of eigenvectors, and L = a diagonal matrix of eigenvalues. *Eigenvectors* are vectors of the factor loadings, which are the weights of each variable on each factor. *Eigenvalues* are an index of the magnitude of each factor and represent the sum of squared factor loadings for a factor.

When EFA is used to evaluate test structure, the analyst must decide how many of the resulting factors are real because some factors may be trivially small. When the hypothesized structure of a test is known, the analyst attempts to identify the hypothesized factors in the solution. If other, meaningful factors emerge, it could signal a lack of validity of the assessment, or it could result in a renewed understanding of the construct.

Confirmatory factor analysis. Given that researchers typically have a theoretical dimensional structure in mind when seeking validity evidence based on test structure, CFA is an attractive alternative to EFA. CFA comes from structural equation modeling, which is a comprehensive procedure for

analyzing the relationships among multiple variables. CFA represents a measurement model within structural equation modeling, in which a model is posited that describes which items are loading on which factors and how the factors are related to one another. The CFA model is

$$y = \Lambda\eta + \varepsilon \tag{4.10}$$

where y is a $(p \times 1)$ column vector of scores for person i on p items, Λ is a $(p \times 1)$ column vector of factor loadings of the p items on the latent factor, η is the latent variable score for person i, and ε is an $(N \times 1)$ column vector of measurement residuals.

If the test data are dichotomous or polytomous, tetrachoric and polychoric correlations and their asymptotic covariances typically serve as input for a CFA. The hypothesized factor loadings, taken from the test specification blueprint, for example, are used to specify the model. Correlations among the factors can also be modeled.

Once a model is specified, it can be fit to the data, and the goodness of fit can be evaluated. Numerous descriptive indices evaluate how well a hypothesized model fits the observed data. Three indices often used in the context of CFA are the root-mean-square error of approximation (RMSEA), the standardized root-mean-square residual (SRMR), and the adjusted goodness-of-fit index (AGFI). The RMSEA is an index of the average residual variance in the data unaccounted for by the model. The SRMR is the standardized difference between the observed covariance and the predictive covariance. For both RMSEA and SRMR, a value of zero indicates perfect fit. The AGFI is an index of the proportion of variance in the data accounted for by the model after adjusting for the number of parameters fit to the data. Rules of thumb for interpreting fit of the model to the data using the standard error of the mean suggest that the RMSEA and SRMR should be less than .10 (and preferably below .05), and the AGFI should be .90 or higher (Browne & Cudek, 1993; Byrne, 1998; Kline, 2005; Mulaik et al., 1989). However, the literature on goodness of fit in structural equation modeling is extensive, and other indices and guidelines have also been recommended.

In addition to testing the fit of specific models of internal structure to observed data, CFA can also be used to evaluate the relative superiority in fit of competing models. For example, if one were interested in determining whether a unidimensional model could be used to represent test performance, the fit of that model could be compared with that of a higher dimensional model. If maximum likelihood is used in fitting the CFA model, the difference in fit between any two nested models can be tested for statistical significance (Raju, Laffitte, & Byrne, 2002).

Multidimensional scaling. MDS can also be used to provide validity data based on internal structure. MDS is a data-analytic procedure that fits dimensions to data so that the underlying structure of the data can be understood. In evaluating the structure of an assessment, either distances are computed among items or an interitem correlation matrix is calculated and then converted to dissimilarities. In single-group analyses, the matrix of observed item dissimilarities is modeled in 1, 2, . . ., or R-dimensional space; the MDS model provides a representation of the observed response data in any R-dimensional item space as

$$d_{jj'} = \sqrt{\sum_{r=1}^{R}(x_{jr} - x_{j'r})^2}. \tag{4.11}$$

Thus, items can be presented by their coordinates, x_{jr} for item j on dimension $r\,(r = 1,...,R)$, or graphically displayed.

The fit of an MDS model to test data is typically evaluated using the fit values STRESS and R^2. *STRESS* represents the square root of the normalized residual variance, and so the smaller the stress value is, the better the fit of the model to the data. R^2 represents the proportion of variance accounted for by the MDS model, and so the larger the value, the better. As with EFA, the interpretability of the solution also plays a major role in determining the MDS solution that best represents the data (and hence test structure).

Analyzing fit of item response theory models. Today, many educational and psychological assessments are developed using IRT. Because IRT posits a specific measurement model, the fit of the model to the data can be directly evaluated to assess internal structure. Although multidimensional IRT models do exist, unidimensional models are much more commonly used. Thus, most IRT residual analyses of

internal structure seek to determine whether (a) the test data are truly unidimensional and (b) the specific IRT model used fits the data.

A variety of IRT models exist, and all describe the probability that an examinee at a particular point on the proficiency continuum (denoted θ) will provide a particular response. A full description of IRT is beyond the scope of this chapter; interested readers are referred to Chapter 6 in this volume for more information.

During IRT model fit analyses, the actual proportions of examinees at various intervals along the θ continuum are plotted along the item characteristic curve that displays the IRT model-based probability that examinees at any point along θ will answer the item correctly. The degree to which these observed proportions deviate from the predicted value represents model misfit.

Evaluating invariance of internal structure. EFA, CFA, MDS, and IRT residual analysis all represent important means for evaluating the internal structure of an assessment and comparing it with the structure hypothesized by the theory of the underlying construct measured. Similar to concerns raised in analyses of criterion-related validity evidence, such as differential predictive validity, the degree to which the internal structure of an assessment is invariant across subgroups of examinees is often an important validity question. CFA and MDS are excellent statistical methods for evaluating the invariance of test structure across multiple groups because the data for multiple groups can be analyzed in a single analysis (Sireci, Patsula, & Hambleton, 2005; Sireci & Wells, 2010). At the item level, invariance can be evaluated by inspecting differential item functioning (DIF).

Differential item functioning. As with the impact at the total test score level discussed earlier with respect to differential predictive validity, impact can also occur at the item level. When such differences occur, they could reflect a true group difference with respect to what the item measures, or the item could be biased in some way to members of one group. The term *differential item functioning* refers to the situation in which examinees who have equal standing on the construct measured by the test but who are from different groups (e.g., ethnicity, sex) have different probabilities of responding to the item (Holland & Thayer, 1988). DIF represents a statistical interaction between group membership and item performance, after matching examinees across groups on some criterion (usually total test score).

DIF, by proxy, indicates multidimensionality and thus represents a difference in a secondary proficiency or item parameter after conditioning on the skill or ability the test was intended to measure (Camilli & Shepard, 1994; Roussos & Stout, 1996). DIF is a necessary but insufficient condition for a claim to be made that an item is biased. For bias to exist, the secondary ability must be an unintended component irrelevant to the purpose of testing. Thus, DIF is determined on the basis of statistics alone, and bias is determined only after follow-up studies prompted by DIF results.

Several methods for assessing DIF exist, including methods based on contingency table analysis such as the Mantel–Haenszel method (Holland & Thayer, 1988) and methods based on IRT such as the likelihood ratio method (Thissen, Steinberg, & Wainer, 1993). A complete description of these methods is beyond the scope of this chapter; interested readers are referred to Holland and Wainer (1993) or Chapter 7 in this volume. Essentially, all DIF procedures match examinees from different groups on some measure of the construct of interest (typically, total test score) and then look for differences in item performance after the matching takes place.

Evaluating the invariance of test structure. As noted at the beginning of this chapter, validity is sometimes described as the degree to which a test measures what it purports to measure. But what if what it measures changes depending on the characteristics of the examinees who take it? The degree to which the construct measured by a test is consistent across subgroups is known as *construct equivalence.* Concerns regarding construct equivalence often arise in cross-cultural research in which constructs such as intelligence or conscientiousness can be culturally dependent (van de Vijver & Poortinga, 2005). It also arises in the case of test accommodations or computer-based testing when the test administration conditions are altered, and the degree

to which such alterations affect the construct is unknown.

There are many aspects of evaluating construct equivalence, both statistical and qualitative. For example, in cross-cultural assessment, a first step is often establishing that the construct is legitimate and appropriate in all cultures assessed. Statistically, test specialists study *structural equivalence,* the degree to which the internal structure of an assessment is consistent, or invariant, across subgroups. Such subgroups could be defined by culture, language, test administration condition, or other factors. (See Volume 3, Chapter 26, this handbook, for considerably more information on equivalence.)

CFA and weighted MDS can both be used to evaluate structural equivalence because they are able to analyze the structure of data from multiple groups simultaneously. In CFA, the degree to which the hypothesized structure of an assessment adequately fits the data for multiple groups can be analyzed using descriptive measures of fit such as RMSEA, SRMR, and AGFI. Alternatively, the hypothesized structure of an assessment can be constrained to be equal across all groups and the fit of that model can be statistically compared with models in which different degrees of constraint are relaxed.

A typical hypothesis tested using CFA is whether the factor loading matrix is equivalent across all groups. Complete structural invariance would result if a model that estimated all parameters separately for each group did not exhibit statistically significant improvement in fit (using the likelihood ratio test) over the model constraining the factor loadings, errors associated with those loadings, and factor correlations to be equivalent across groups. When numerous groups are involved in the analyses, descriptive fit indices are more generally used (Byrne & van de Vijver, 2010).

Evaluating structural invariance using MDS requires weighted MDS, in which a weight is incorporated into the distance model to adjust the overall structure of the assessment to best fit the data for each group. Specifically,

$$d_{ijk} = \sqrt{\sum_{a=1}^{r} w_{ka}\left(x_{ia} - x_{ja}\right)^2}, \qquad (4.12)$$

where d_{ijk} is equal to the Euclidean distance between stimuli (e.g., test items) i and j for group k, w_{ka} is the weight for group k on dimension a, x_{ia} is the coordinate of stimulus i on dimension a, and r is the dimensionality of the model. The weights for each subgroup on each dimension (w_{ka}) contain information regarding structural equivalence. If the pattern of weights is similar across groups, the structure is consistent across the groups. If a dimension is needed to account for variation among the items in one group but not another, that group will have a relatively large weight on that dimension, whereas the other groups will have much smaller weights on the dimension. Such a finding indicates a lack of structural equivalence.

CFA and weighted MDS can be used to evaluate test structure across groups in complementary fashion (Sireci & Wells, 2010). CFA evaluates the hypothesized test structure, whereas MDS is an exploratory analysis that fits dimensions to best account for the data in all groups.

Validity Evidence Based on Response Processes

Gathering validity evidence based on response processes is perhaps the most difficult validity evidence to gather because it involves demonstrating that examinees are invoking the hypothesized constructs the test is designed to measure in responding to test items. As the *Standards* (AERA et al., 1999) describe, "Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (p. 12). Gathering this type of evidence is difficult because one cannot directly observe the cognitive processes going on within people's heads as they respond to test items. Although some studies have used MRI to see which regions of the brain are activated when responding to tasks (e.g., Owen, Borowsky, & Sarty, 2004), most studies of response processes use indirect means such as cognitive interviews, think-aloud protocols, focus groups, or analysis of answer patterns and item response time data.

Methods for gathering response process data. Messick (1989) pointed out that the information-processing models in cognitive psychology provide

several means for investigating the response processes used by examinees in responding to items. Gathering evidence based on response processes involves determining the cognitive strategies used by test takers or ruling out specific construct-irrelevant strategies such as guessing or test wiseness. The evidence can take many forms, including interviewing test takers about their responses to test questions, systematically observing test response behavior, evaluating the criteria used by judges when scoring performance tasks, analyzing item response time data, and evaluating the reasoning processes examinees use when solving test items.

Think-aloud protocols and cognitive interviews. *Protocol analysis* refers to both think-aloud protocols and cognitive interviews. In think-aloud protocols, examinees explain what they are thinking as they respond to test items, and their explanations are recorded. Retrospective analyses can also be conducted in which examinees explain why they responded to an item in a particular way. In cognitive interviewing, a specific interview protocol is designed and examinees are asked questions to test whether they are responding to the item in the manner intended (Beatty, 2004).

Think-aloud and interview protocols can be used to see whether examinees are guessing, eliminating distractors in multiple-choice items, or using the hypothesized cognitive strategies intended by the test developers. An advantage of the cognitive interview is that specific hypotheses can be tested. For example, if there is a construct-irrelevant threat that may affect test performance, such as test wiseness, the interview can include questions about such threats. Cognitive interviews and think-aloud protocols can also test specific theories regarding why an item may exhibit DIF across groups (e.g., Ercikan et al., 2010).

Chronometric analysis. *Chronmetric analysis* refers to the analysis of item response time data, which record how long it takes an examinee to respond to a test item. Before the age of computer-based testing, gathering such data was hard, but doing so is relatively easy when a test is administered on a computer. One analysis that could be done is to test whether examinees take more time to answer items that are hypothesized to require greater processing load. If so, support for the differentiation of the specified cognitive levels measured by the test is provided.

Wise and his colleagues (Wise, 2006; Wise & Kong, 2005) have used the amount of time examinees take to respond to items to measure rapid guessing behavior and the degree to which examinees are engaged in the test. Identifying such unmotivated responses helps explain the performance of certain examinees and can be used to identify unmotivated examinees and adjust group statistics when a lack of motivation attenuates aggregate statistics.

Evidence-centered test design. Many test specialists have argued that the cognitive skills measured on educational tests need to be more carefully specified using cognitive theories (Huff, Steinberg, & Matts, 2010; Mislevy, 2009; Snow & Lohman, 1989). The incorporation of cognitive theories into test development is often referred to as *evidence-centered design* or *principled assessment development*. The general idea underlying this approach is that the skill or objective measured by each test item is clearly specified in terms of the skill the item measures and the different types of responses that indicate different levels of the skill. Different models can be established, such as the task model that specifies the claims regarding what the item measures and the observable evidence in the form of examinee responses. Mathematical models can be developed that relate the different sets of skills measured by the test and which ones are needed to succeed on specific test items. The degree to which success on the test and item difficulty are congruent with the specification of the skills measured provides evidence that the targeted cognitive processes are being measured and hence supports valid interpretations of the test scores.

Mathematical modeling of item difficulty. Similar to evidence-centered design, some educational test specialists have investigated modeling item difficulty on the basis of specific attributes of test items related to cognitive processes (Mulholland, Pellegrino, & Glaser, 1980; Sheehan & Mislevy, 1990; Tatsuoka, 1987; Whitely, 1983). Many of these models are extensions of IRT in which different item attributes

are treated as facets. In other cases, the difficulty is modeled after the test is assembled to provide more diagnostic information regarding examinees' performance (Sheehan, 1997). When item difficulty can be explained using the cognitive attributes of the items, validity evidence based on response processes is provided.

Evaluating processes used by graders. One other area of evidence based on response processes focuses not on examinees but on the people who score their responses. Whenever tests involve graders or observers, such as when students write essays that are scored by graders or when observers judge the performance of an examinee performing a task, the degree to which the graders rate performance in accordance with the scoring rubric is a critical validity issue. Analyses of grader accuracy, scoring rubric stability, and interrater reliability all provide important validity evidence based on the response processes of the graders.

Other evidence based on response processes. In addition to the methods mentioned thus far, several other methods for gathering validity evidence based on response processes have been proposed. These methods include analysis of eye movements in which the direction and duration of examinees' eye movements during task performance are measured (Messick, 1989), analysis of systematic errors such as performing an analysis of the incorrect response options on a multiple-choice test (Abedi, 2007; Thissen, Steinberg, & Fitzpatrick, 1989), and analysis of omit rates. When items are left blank by many examinees, reasons for such omit rates should be explored because they are likely to influence score interpretations. Another method used to evaluate response processes is review of scratch paper and other draft material students create in responding to items.

Summary of validity evidence based on response processes. Validity evidence based on response processes is typically hard to gather, and so examples of comprehensive studies in this area are rare. However, computer-based testing, evidence-centered design, and other developments offer promise for more research in this area. Regardless of the method used, the quality of the data gathered must be considered, particularly when such data are based on subjective judgments such as observers and interviewers. Biases in examinees' responses to observations and interviews, such as social desirability, must also be considered. Although such potential problems exist in gathering response process data, the effort put into gathering such data is typically well worth it, because data based on response processes represent a unique perspective from which test score interpretations can be evaluated.

Validity Evidence Based on Consequences of Testing

Validity evidence based on consequences of testing refers to evaluating both the intended and the unintended consequences associated with a testing program. Tests are used to promote positive consequences such as appropriate diagnosis of psychological disorders, protection of the public, improved instruction, and better understanding of the constructs measured. Unintended positive consequences that were not explicitly intended or envisioned may also emerge. However, unintended negative consequences may also occur in a testing program. Examples of unintended consequences may be adverse impact that leads to decreased education and employment opportunities for members of certain groups, increased dropout rates in schools, and poor decisions regarding resource allocations or employees' salaries on the basis of test performance.

Whether validity evidence based on the consequences of testing is relevant in evaluating the validity of inferences derived from test scores is a subject of some controversy (Popham, 1997; Shepard, 1997). Considerations of testing consequences are an important social policy issue, but many test specialists believe they are extraneous to validity. However, others see the evaluation of testing consequences as a critical element in evaluating the appropriateness of using a test for a particular purpose (e.g., Messick, 1989). We believe this debate is one of nomenclature, and given that virtually all testing programs have consequences on some level, it is important to evaluate the degree to which the positive outcomes of the test outweigh any negative consequences.

Gathering validity evidence based on consequences. Gathering validity evidence based on testing consequences should start with an analysis of potential consequences. Identification of the positive consequences starts with the stated purposes of the test. Is the test fulfilling its purposes? Evidence to the affirmative should support claims of positive consequences. Identification of negative consequences can be easy in the context of high-profile or high-stakes tests because concerned citizens or special interest groups often loudly criticize such tests. These criticisms are fertile soil for identifying potential negative consequences. For example, critics of state-mandated educational tests often claim that these tests take away valuable instructional time and narrow the curriculum. An analysis of how tests have affected instruction could investigate both the positive claim that the tests are improving instruction (e.g., by providing valuable information to teachers) and the negative claim that the tests are narrowing the curriculum.

Identifying the consequences to study depends entirely on the testing purpose and context. In employment testing, for example, adverse impact (as when hiring or promotion decisions that are based on test scores lead to lower acceptance rates for underrepresented minority examinees than for non-minority examinees) is important to study because an unnecessarily less diverse workforce is an injustice to society, and not getting a job or a promotion is a dire consequence for an individual. In clinical assessment, improper diagnosis of a disorder may lead to unhelpful or even harmful treatment. In educational testing, differential referral rates for remediation may be a potential concern. Similarly, when tests are used to evaluate instructional programs, such as native language instruction, test scores may serve as one impetus for closing some programs or getting rid of such instruction altogether. Whenever such negative consequences occur, the technical quality of the test must be demonstrated, as should the benefits associated with the testing program.

A valuable way to gather evidence of testing consequences is to gather feedback from test takers and other stakeholders (e.g., clinicians, teachers, policymakers). Surveys, interviews, and focus groups can be used to gather data on consequences from these groups. Evaluating trends in test performance over time, such as tracking student achievement, graduation rates, or diagnostic classifications over time (and across subgroups), are other important means. In the context of educational achievement testing, other examples of evidence based on testing consequences that could be used to support the use of a test could be analysis of educational gains associated with testing programs, the degree to which the tests have positively influenced instruction and provided professional development for teachers (Cizek, 2001), the effects of the test on retention and dropout, and the degree to which the tests may have increased parents' involvement in their children's education.

Validity evidence based on testing consequences has received a great deal of attention in the courtroom. For example, adverse impact on educational tests, employment tests, and licensure tests has led to legal challenges to test use. In fact, an analysis of testing consequences is one way a test can be legally challenged owing to (a) Title VI of the Civil Rights Act of 1964, (b) Title VII of the Civil Rights Act of 1964, and (c) the Equal Protection Clause of the 14th Amendment (Sireci & Parker, 2006). Essentially, these laws allow plaintiffs to challenge the appropriateness of a test for such high-stakes decisions whenever disparate impact occurs. For this reason, these testing programs monitor the success rates for various subgroups of examinees. If a test is challenged in court for the unintended consequence of disparate impact, other validity evidence is used to defend the test, mainly evidence based on test content and appropriate test construction and standard setting practices (Sireci & Green, 2000; Sireci & Parker, 2006).

Who should gather validity evidence based on testing consequences? Given the importance of studying the consequences of testing, an important question is, "Who should do it?" Likely candidates are test developers (e.g., a test development contractor), testing agencies (e.g., a state department of education), and test users (e.g., a school district using a commercial test). Virtually all partners in the testing process, including measurement professionals, could, and probably should, be involved.

However, studying testing consequences is quite a bit harder than computing a coefficient alpha or a Mantel–Haenszel statistic. Thus, although the endeavor is important, it is likely to be expensive and time consuming, with few volunteers stepping forward to do it.

Nevertheless, the responsibility remains, and there are good examples of comprehensive assessments of testing consequences (e.g., Lane, Parke, & Stone, 1998; Taleporos, 1998). Test developers are responsible for providing evidence that a test measures what it claims to (evidence based on test content), and according to the *Standards*, they are also responsible for warning against inappropriate test use (AERA et al., 1999, pp. 17–18). However, test developers cannot be expected to conduct longitudinal studies of tests they develop for one purpose that are ultimately used for another, unanticipated purpose. For this reason, development of clear and comprehensive statements of the purpose and intent of a testing program is critical.

Messick (1989), Popham (1997), and Shepard (1993, 1997) argued that much of the responsibility for investigating the consequences of test use lies with test users. Messick referred to this as an ethical responsibility of test users, because they are in the best position to evaluate the value implications specific to their setting. In some instances, test users adapt existing tests for a new purpose. In these circumstances, the test user is clearly responsible for evaluating testing consequences. Although gathering such evidence may be difficult, as are all validation endeavors, determining the consequences to be studied and devising ways to gather and analyze the data requires creativity on the part of the validator.

SUMMARY

In this chapter, we have provided a historical view of validity theory and have described current conceptualizations of validity. In addition, we have described the different types of evidence that can be gathered to evaluate the appropriateness of the use of a test for a particular purpose. Our categorization of the evidence has followed the current version of the *Standards for Educational and Psychological Testing* (AERA et al., 1999). We also have described common statistical techniques used to analyze several types of validation data. Our hope is that these discussions and descriptions advance understanding of how tests should be developed and evaluated. We also hope our descriptions of the statistical procedures and our cautions regarding conducting such analyses empower researchers to conduct more comprehensive and informative validation studies. The numerous references to theoretical and applied validity research we provided in this chapter should be valuable resources to those interested in learning more about validity theory and test validation.

One point that should be clear throughout this chapter is that any serious validation effort must be comprehensive and involve multiple sources of evidence that bear on the degree to which a test sufficiently fulfills its purpose. The evidence should be integrated with theory and focus on (a) the degree to which test score interpretations are congruent with their intended purposes and (b) the defensibility of those interpretations. The evidence should reference theories underlying the construct measured and how the test development and evaluation processes are true to measurement of that construct. If such an integrated validity argument is developed, the strengths and limitations of the use of a test for a particular purpose will be documented, and better decisions regarding the use of tests will be made.

References

Abedi, J. (2007). English language learners with disabilities. In C. C. Laitusis & L. Cook (Eds.), *Large scale assessment and accommodations: What works?* (pp. 53–65). Arlington, VA: Council for Exceptional Children.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51,* 1–38.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.

American Psychological Association. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

American Psychological Association Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist, 7*, 461–475. doi:10.1037/h0056631

Beatty, P. (2004). The dynamics of cognitive interviewing. In S. Presser, J. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questions* (pp. 45–66). Mahwah, NJ: Wiley.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9). Retrieved from http://escholarship.bc.edu/jtla/vol6/9

Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique, 12*, 191–244.

Binet, A., & Henri, B. (1899). La psychologie individuelle [The psychology of the individual]. *L'Année Psychologique*, *2*, 411–465.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York, NY: Harper.

Browne, M. W., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 445–455). Newbury Park, CA: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Hillsdale, NJ: Erlbaum.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing issues of non-equivalence. *International Journal of Testing, 10*, 107–132. doi:10.1080/15305051003637306

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi:10.1037/h0046016

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*, 19–27. doi:10.1111/j.1745-3992.2001.tb00072.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*, 179–194. doi:10.1207/s15324818ame0202_6

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi:10.1037/h0040957

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Ebel, R. L. (1956). Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement, 16*, 269–282. doi:10.1177/001316445601600301

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*, 640–647. doi:10.1037/h0045478

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*, 24–35. doi:10.1111/j.1745-3992.2010.00173.x

Garrett, H. E. (1937). *Statistics in psychology and education*. New York, NY: Longmans, Green.

Glasnapp, D., Poggio, J., Carvajal-Espinoza, J., & Poggio, A. (2009, April). *More evidence: Computer vs. paper and pencil delivered test comparability*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427–438.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hox, J. J. (1995). *Applied multi-level analysis* (2nd ed.). Amsterdam, the Netherlands: TT-Publikaties.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied*

Measurement in Education, 23, 310–324. doi:10.1080/08957347.2010.510956

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology, 10*, 93–98. doi:10.1037/h0059212

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: Rowman & Littlefield.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. doi:10.1037/0033-2909.112.3.527

Kelley, T. L. (1927). *Interpretation of educational measurement.* Yonkers-on-Hudson, NY: World Book.

Kim, D., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *Journal of Technology, Learning, and Assessment, 6*(4). Retrieved from http://escholarship.bc.edu/jtla/vol6/4

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Korbin, J. (2010, April). *Exploring the variability in the validity of SAT scores for predicting first-year college grades at different colleges and universities.* Paper presented at the meeting of American Educational Research Association, Denver, CO.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*, 24–28. doi:10.1111/j.1745-3992.1998.tb00830.x

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294–304. doi:10.1177/001316445601600303

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(Monograph Suppl. 9), 635–694.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment among curriculum, assessments, and instruction. *Review of Educational Research, 79*, 1332–1361. doi:10.3102/0034654309341375

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Phoenix, AZ: Oryx Press.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 83–108). Charlotte, NC: Information Age.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness of fit indices for structural equation models. *Psychological Bulletin, 105*, 430–445. doi:10.1037/0033-2909.105.3.430

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*, 252–284. doi:10.1016/0010-0285(80)90011-0

Owen, W. J., Borowsky, R., & Sarty, G. E. (2004). fMRI of two measures of phonological processing in visual word recognition: Ecological validity matters. *Brain and Language, 90*, 40–46. doi:10.1016/S0093-934X(03)00418-8

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III: Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A, 187*, 253–318. doi:10.1098/rsta.1896.0007

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 2*, 559–572.

Popham, W. J. (1997). Consequential validity: Right concern–wrong concept. *Educational Measurement: Issues and Practice, 16*, 9–13. doi:10.1111/j.1745-3992.1997.tb00586.x

Porter, A. C., Smithson, J. L., Blank, R. K., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*, 27–51.

Pressey, S. L. (1920). Suggestions looking toward a fundamental revision of current statistical procedure, as applied to tests. *Psychological Review, 27*, 466–472. doi:10.1037/h0075018

Puhan, G., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment, 6*(3). Retrieved from http://escholarship.bc.edu/jtla/vol6/3

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517–529. doi:10.1037/0021-9010.87.3.517

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification exams. *Applied Measurement in Education, 14*, 369–415. doi:10.1207/S15324818AME1404_4

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF paradigm. *Applied Psychological Measurement, 20*, 355–371. doi:10.1177/014662169602000404

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 290–296.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and

employment. *American Psychologist, 63*, 215–227. doi:10.1037/0003-066X.63.4.215

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*, 112–118. doi:10.1037/0021-9010.85.1.112

Schmidt, F. L. (1988). Validity generalization and the future of criterion-related validity. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 173–189). Hillsdale, NJ: Erlbaum.

Sheehan, K. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333–352. doi:10.1111/j.1745-3984.1997.tb00522.x

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*, 255–272. doi:10.1111/j.1745-3984.1990.tb00747.x

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5–24. doi:10.1111/j.1745-3992.1997.tb00585.x

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45*, 83–117. doi:10.1023/A:1006985528729

Sireci, S. G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299–321. doi:10.1207/s15326977ea0504_2

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age.

Sireci, S. G., & Green, P. C. (2000). Legal and psychometric criteria for evaluating teacher certification tests. *Educational Measurement: Issues and Practice, 19*, 22–31, 34. doi:10.1111/j.1745-3992.2000.tb00019.x

Sireci, S. G., & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: The sensitivity review process. *CLEAR Exam Review, 5*(2), 22–28.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice, 25*, 27–34. doi:10.1111/j.1745-3992.2006.00065.x

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptations process. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Hillsdale, NJ: Erlbaum.

Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457–490. doi:10.3102/00346543075004457

Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admissions Test scores. *Educational and Psychological Measurement, 66*, 305–317. doi:10.1177/0013164405282455

Sireci, S. G., & Wells, C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33–68). Washington, DC: Council of Chief State School Officers.

Smith, H. L., & Wright, W. W. (1928). *Tests and measurements.* New York, NY: Silver, Burdett.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). Phoenix, AZ: Oryx Press.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 15*, 201–293. doi:10.2307/1412107

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and Practice, 17*, 20–23. doi:10.1111/j.1745-3992.1998.tb00829.x

Tatsuoka, K. K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24*, 233–245. doi:10.1111/j.1745-3984.1987.tb00277.x

Teng, S.-Y. (1943). Chinese influence on the Western examination system. *Harvard Journal of Asiatic Studies, 7*, 267–312. doi:10.2307/2717830

Terman, L. M., & Childs, H. G. (1912). A tentative revision and extension of the Binet-Simon measuring scale of intelligence. *Journal of Educational Psychology, 3*, 61–74. doi:10.1037/h0075624

Terman, L. M., Lyman, G., Ordahl, G., Ordahl, L., Galbreath, N., & Talbert, W. (1915). The Stanford revision of the Binet-Simon scale and some results from its application to 1000 non-selected children. *Journal of Educational Psychology, 6*, 551–562. doi:10.1037/h0075455

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161–176. doi:10.1111/j.1745-3984.1989.tb00326.x

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Thomas, R. D., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research, 45*, 253–275. doi:10.1023/A:1006954016433

Thorndike, E. L. (1931). *Measurement of intelligence*. New York, NY: Bureau of Publishers, Columbia University.

Thurstone, L. L. (1932). *The reliability and validity of tests*. Ann Arbor, MI: Edwards.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Hillsdale, NJ: Erlbaum.

Wainer, H., & Sireci, S. G. (2005). Item and test bias. In *Encyclopedia of social measurement* (Vol. 2, pp. 365–371). San Diego, CA: Elsevier. doi:10.1016/B0-12-369398-5/00446-1

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Washington, DC: Council of Chief State Schools Officers.

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197. doi:10.1037/0033-2909.93.1.179

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*, 95–114. doi:10.1207/s15324818ame1902_2

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183. doi:10.1207/s15324818ame1802_2

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*, 10–14.

CHAPTER 5

FACTOR ANALYSIS OF TESTS AND ITEMS

Li Cai

When broadly conceived, *factor analysis* can be defined as a body of interrelated statistical and psychometric techniques that are particularly useful for exploring and testing the structure of psychological assessment tools. Factor analysis seeks to uncover the relationships between the observed test scores (or item responses) and the hypothesized latent variables (factors) that represent the psychological constructs being measured as well as the associations among the latent variables themselves. On one hand, the latent factors may be posited as statistical devices that can explain the observed individual differences in test scores. On the other hand, the number, nature, and interrelatedness of factors as seen through the lens of statistical factor analysis may lend empirical support to the reliability and validity of psychological assessments. Before discussing factor analysis methods and all the terminology in detail, however, it is helpful to briefly review the history behind this widely used (and abused) method.

HISTORICAL BACKGROUND

Factor analysis made its debut when Spearman (1904) published his seminal article on general intelligence. Spearman produced the world's first table of factor loadings, presented in the form of correlations of the observed test scores (e.g., classics, English, various sensory discrimination tasks) with the latent general intelligence factor (*g*) presumed to be the underlying cause of the observed pattern of correlations (p. 276). The ingenious methods that he used to obtain those numbers are based on partial correlations. Thus, the early phase of factor analysis was as much about the analysis of correlation matrices as about the theory of primary mental abilities. Somewhat unfortunately, the early history of factor analysis as a statistical method was overshadowed by a debate about the existence and heritability of Spearman's *g*. A recent discussion about these early developments can be found in Cudeck and MacCallum's (2007) excellent edited volume (particularly the first 4 chapters).

In the 1930s and 1940s, L. L. Thurstone initiated major new developments in factor analysis at the University of Chicago (Thurstone, 1947). He clearly described the statistical factor analysis model as a linear model that was based on his factor-analytic theory of primary mental abilities. His factor analysis model is now known as the *common factor model*. The model implies a certain covariance structure model, which can be fitted to a sample of data and falsified via model fit tests. This conceptualization of factor analysis has since become dominant in practice, and statisticians have developed comprehensive and elegant theories for the analysis of linear covariance structures (see, e.g., Browne & Arminger, 1995). In particular, normal theory maximum likelihood, pioneered by Lawley and Maxwell (1963) and

This research was supported by Institute of Education Sciences Grants R305B080016 and R305D100039 and National Institute on Drug Abuse Grants R01DA030466 and R01DA026943. The views expressed in this chapter do not reflect the views and policies of the funding agencies.

DOI: 10.1037/14047-005
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

subsequently extended and perfected by Jöreskog (1969), now provides a solid statistical basis for factor analysis.

Applications of factor analysis of dichotomous item responses began to emerge not long after Thurstone's (1947) development of common factor analysis (Guilford, 1941). However, the common factor model, being a linear model, is not adequate when directly applied to Pearson correlations computed from dichotomous test items that have endorsement probabilities substantially different from one half, resulting in the so-called "difficulty" factors (Ferguson, 1941). By the early 1950s, a new kind of mental test theory that was based on individual item responses began to emerge (e.g., Lazarsfeld, 1950). The ensuing progress (e.g., Lord & Novick, 1968) led to nonlinear statistical models and psychometric methods known collectively as item response theory (IRT). The need to factor analyze item-level data initiated the development of modern item factor analysis methods that are closely related to IRT. Ultimately, an interest in likelihood-based inference for IRT and item factor-analytic models propelled the contemporary view of factor analysis as a member of a large family of generalized linear and latent variable models that has been embraced not only by psychometricians but also by statisticians (Bartholomew & Knott, 1999), especially as Bayesian formulations of hierarchical models have become increasingly popular in the past decade (e.g., Dunson, 2000).

RECURRING ISSUES

As mentioned, factor analysis can be applied to batteries of tests (using the linear common factor analysis model) or it can be applied to test items directly (using the nonlinear item factor analysis model). There are many book-length treatments of factor analysis of tests (e.g., Gorsuch, 1983; Harman, 1976; McDonald, 1985). Recently, some nontechnical review articles on item factor analysis have also appeared in the psychological literature (e.g., Wirth & Edwards, 2007). Although the underlying linear factor analysis model and the nonlinear item factor analysis model have important mathematical differences, a number of similar methodological issues arise in the applications of both. They are briefly described here, and subsequent sections of this chapter revisit many of them in more detail.

Component Analysis Versus Factor Analysis

The key distinction between principal component analysis and factor analysis is clear. Principal component analysis does not require the use of latent variables in the model, but factor analysis must involve latent variables. Because the computations involved in principal component and factor analysis can be quite similar (or they can be quite different, as in the case of item factor analysis), the two popular multivariate techniques typically reside in the same procedure in many statistical packages (e.g., SPSS), which has contributed to the confusion between the two. As a data condensation or reduction technique, principal component analysis holds unique advantages over factor analysis. However, because it does not involve any latent variables, it is not a formally testable model, unlike the factor analysis model, which is falsifiable.

Exploratory Versus Confirmatory Use

As the names suggest, factor analysis can be either exploratory or confirmatory. When it is used in the exploratory mode, the goal is to extract a number of factors that would result in substantively interpretable factor patterns, adequate model fit, and consistently replicable findings across samples. Under exploratory factor analysis (EFA), the best-fitting factor pattern is suggested by the data, which is in contrast to confirmatory factor analysis (CFA), in which the researcher must suggest a factor pattern before the analysis can begin. Generally speaking, when little is known about a particular domain, factor analysts should typically start in an exploratory mode, but as evidence begins to emerge and accumulate over time, the focus naturally shifts to testing and confirming theoretically driven factor structures, implying an increased reliance on CFA. However, for many studies aimed at developing, improving, or studying the features and psychometric properties of psychological assessment instruments, both EFA and CFA may be useful within the same study. The CFA is typically conducted using a

separate replication sample that is distinct from the initial EFA sample to serve as a form of cross-validation (e.g., Hawkley, Browne, & Cacioppo, 2005).

Dimensionality

In both EFA and CFA, dimensionality may be taken to mean three separate but related identities. At its simplest, *dimensionality* may refer to the number of latent factors in a particular factor analysis model. In this case, the dimensionality of the model is a number that is specified by the researcher. Determining this number is one of the major tasks in EFA. At a more qualitative level, dimensionality may be attached to the psychological assessment instrument itself, as a description of the number of distinct constructs or domains that the instrument intends to measure. Finally, a mathematical definition of dimensionality follows from conditional covariance theory (e.g., Zhang & Stout, 1999). Loosely speaking, a set of psychological test items or a battery of psychological tests is said to be d dimensional if d is the minimal number of latent factors such that after controlling for (or conditioning on) these latent factors, the test items or tests become independent. As such, conditional independence is the hallmark feature of factor analysis models. In practice, approximate conditional independence (e.g., as judged from residual covariance) is an indication of adequate dimensionality specification.

Model Fit Assessment

Whenever a statistical model is fitted to data, the adequacy of fit must be examined. A factor analysis model is no exception. In both EFA and CFA, model fit assessment is strongly related to dimensionality as well as to the distributional specifications. When the number of factors is not correctly specified (over- or underfactoring), the interpretability of the factor solution may be severely undermined. It is a standard assumption in factor analysis of tests and items that the latent factors are multivariate normality distributed. Although this assumption may be adequate for cognitive abilities in the general population, it may be inappropriate when the latent variables correspond to mental illness or special, mixed populations. A variety of statistical tests, fit indices, and residuals are available (and should be examined) to assess the fit of the factor analysis model.

Indeterminacies

Factor analysis models have a number of indeterminacies. The latent variables do not have a priori defined location and scale parameters. The EFA model is also rotationally indeterminate. That is, one can transform the factor solution by arbitrarily rotating the orientation of the axes of the coordinate system of the factor space without altering the fit of the model. Therefore, the initial (unrotated) extraction in EFA should never be directly interpreted. In EFA, many analytical rotation methods exist, with the goal of simplifying the interpretation of the obtained factor solution. Even in the case of (the relatively more stable) CFA model, the factor pattern is still indeterminate under reflection (permutation of signs), that is, one can arbitrarily change the direction of the factors without changing the model fit. More discussions on rotation are provided in a subsequent section. A straightforward explanation (not the only explanation) for the multitude of indeterminacies is the fact that there are many more parameters in the factor analysis model than there are pieces of available observed information. In the absence of additional restrictions called *identification conditions,* there exist an infinite number of solutions that fit the data equally well. Latent variable models, factor analysis models included, all possess indeterminacies.

Hierarchical Versus Higher Order Solutions

Of historical and practical interest is the distinction between hierarchical and higher order factor solutions. The two kinds of solutions are best illustrated with path diagrams—graphical depictions of the model structure particularly useful in factor analysis. As the conventions of path diagrams go, rectangles represent observed variables and circles represent latent variables. Single-headed arrows indicate directional influence.

Figure 5.1 shows a hierarchical factor model for nine observed variables. The factor F represents the target trait being measured, and the specific factors S_1 to S_3 represent residual correlations specific to subsets of tests that are not fully represented by the general dimension F. Historically, one may adopt the methods of Schmid and Leiman (1957) or

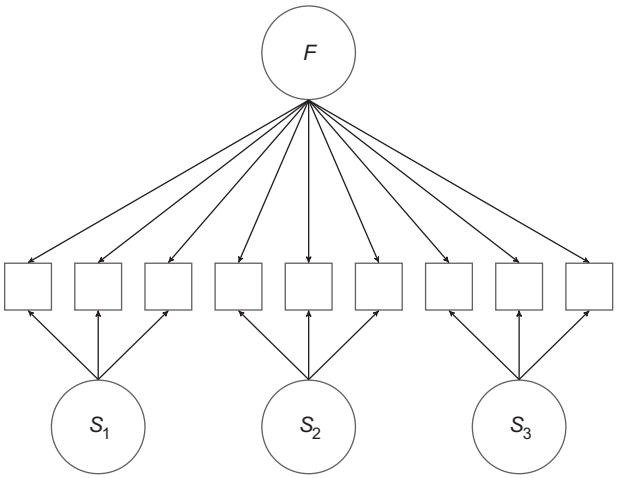

FIGURE 5.1. A hierarchical factor model. F = general factor; S_1, S_2, and S_3 = specific factors.

Wherry (1959) to produce a hierarchical factor solution. The advent of CFA methods made fitting a hierarchical factor model considerably more streamlined. Hierarchical factor solutions are characterized by mutually uncorrelated layers of factors. For a typical hierarchical factor model, each observed variable is directly influenced by exactly one latent variable in a given layer. For instance, there are two layers of factors in Figure 5.1. The specific factors S_1 to S_3 may be called Layer 1 factors because they are more numerous. The general factor F is a Layer 2 factor. When there are exactly two layers, the model is called a *bifactor model* (Holzinger & Swineford, 1937). Efficient methods for item bifactor analysis also exist (Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992).

A higher order factor model is shown in Figure 5.2 for the same tests as in Figure 5.1. Factors F_1 to F_3 are called *first-order factors.* They are influenced by (regressed on) a single second-order factor G. In this model, G is still presumed to represent the target trait. A feature of the higher order factor model is that the observed variables receive only indirect influence from the second-order factor G. All direct influences must come from the first-order factors. In an important article, Yung, McLeod, and Thissen (1999) revealed an interesting relation between the higher order factor model and the hierarchical factor model. They showed analytically that the class of higher order factor models is in fact nested within the class of hierarchical factor models, which

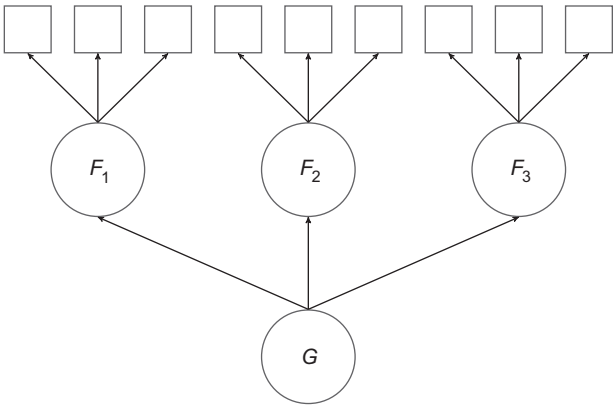

FIGURE 5.2. A higher order factor model. F_1, F_2, and F_3 = first-order factors; G = second-order factor.

implies that one can quantitatively test the tenability of the higher order model. F. F. Chen, West, and Sousa (2006) empirically compared the higher order and hierarchical factor models for quality-of-life data and found that the hierarchical (in their case, bifactor) model tends to fit data better.

FACTOR ANALYSIS OF TESTS

The descriptions of factor analysis of tests are largely based on the classical linear common factor model. These methods assume that the test scores are continuous. Throughout the rest of this chapter, the assumption is that the reader has some familiarity with matrices (at least at a superficial level), which are rectangular arrays of numbers.

Common Factor Model

Consider a data matrix $\mathbf{Y}$ with N rows (each row is for an individual) and n columns (each column is for a variable). In other words, the ijth element of this matrix, denoted as y_{ij}, is the score of person i on test or variable j. The common factor model specifies the following linear regression equation for the observed test score:

$$\begin{aligned} y_{ij} &= \mu_j + \lambda_{j1}\eta_{i1} + \cdots + \lambda_{jp}\eta_{ip} + \varepsilon_{ij} \\ &= \mu_j + \sum_{k=1}^{p} \lambda_{jk}\eta_{ik} + \varepsilon_{ij}, \end{aligned} \tag{5.1}$$

where μ_j is the mean of observed variable j, η_{ik} the ith person's score on common factor k, and ε_{ij} is the ith person's score on the jth unique factor. The

regression coefficient λ_{jk} is referred to as the factor loading of test j on common factor k. Because the loadings are regression coefficients, they can be interpreted in the same manner as the expected increase in y given a 1-unit increase in η, holding other ηs constant. When both y and η are standardized, which is routinely done in EFA, the loadings become standardized regression coefficients. In brief, a common factor model regresses the observed test scores (outcome variables) on the latent factor scores (predictor variables). Equation 5.1 reveals the fundamental difficulty of factor analysis: The predictors are completely unobserved.

The ηs in Equation 5.1 are the primary abilities, traits, and propensities that are presumed to influence more than one test. They are referred to as common factors because they are common to more than one observed variable, and as such the individual differences on these common factors induce correlations among the observed test scores. The εs are specific to only one test, and they do not explain correlations among the observed variables. Hence, they are referred to as the *unique factors*. The unique factor is made up of two parts: systematic and error of measurement. In the linear common factor model, there are altogether $p + n$ factors, p the common and n the unique. The factor loadings indicate the direction and magnitude of the influence of common factors on observed test scores, where 0 indicates no influence. One can infer the nature of the factors from the pattern of loadings. If a subset of observed variables is substantially influenced by one common factor, the shared characteristics of this subset of variables provide a basis for naming and interpreting the meaning of the factor (or the lack of meaning). Thus, a central goal of statistical factor analysis is to obtain accurate estimates of factor loadings.

Some derived parameters are also of substantive interest. For example, the proportion of variance in the jth observed variable that is explained by the common factors is referred to as *communality*. This value is analogous to the squared multiple correlation coefficient routinely reported in regression analysis. One minus communality is the *uniqueness*, that is, the proportion of variance in an observed variable that is due to the associated unique factor. This value shows that the observed variance is decomposed by the factor analysis model into two parts: common variance and unique variance.

With some help from matrix algebra, the common factor model for a vector of scores from the same person can be represented in a matrix equation. Let $\mathbf{y}_i$ be an $n \times 1$ column vector of scores from respondent i. Similarly, let $\boldsymbol{\eta}_i$ be a $p \times 1$ vector of common factor scores and $\boldsymbol{e}_i$ an $n \times 1$ vector of unique factor scores. Let $\boldsymbol{\mu}$ be an $n \times 1$ mean vector. Then the common factor model is

$$\begin{pmatrix} y_{i1} \\ \vdots \\ y_{in} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1p} \\ \vdots & \vdots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{np} \end{pmatrix} \begin{pmatrix} \eta_{i1} \\ \vdots \\ \eta_{ip} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in} \end{pmatrix},$$

or, even more compactly,

$$\mathbf{y}_i = \boldsymbol{\mu} + \Lambda\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{5.2}$$

where the $n \times p$ matrix of coefficients Λ is known as the *factor loading matrix*. The jth row of this matrix contains the loadings of manifest variable j on all p latent common factors. The kth column of this matrix contains the loadings of all observed variables on common factor k. Equation 5.2 is often referred to as the *factor analysis data model*.

Derived Mean and Covariance Structure Model

From the data model, a mean and covariance structure model can be derived. Two basic assumptions are required. First, both the common factors and the unique factors have zero means. Second, the common factors and the unique factors are assumed to be uncorrelated. Translated into matrices, the first assumption implies that the expected values of the observed test scores are $E(\mathbf{y}_i) = \boldsymbol{\mu}$. In other words, the factor analysis model implies a saturated mean structure; that is, the model does not impose additional restrictions on the observed variables' means. Other than simultaneous factor analysis in several populations (Sörbom, 1974), one may ignore the mean structure for practical purposes. Indeed, analysis of correlation matrices also eliminates the mean structure altogether. Let the covariance matrix of the common factors be Φ and that of

the unique factors be Δ. The implied factor analysis covariance structure model is

$$\text{var}(\mathbf{y}_i) = \Sigma = \Lambda\Phi\Lambda^t + \Delta. \quad (5.3)$$

Equation 5.3 is the fundamental factor analysis covariance structure model.

A Numerical Example

These technical issues are best illustrated with an example. Consider a hypothetical data set consisting of four tests. Suppose the first two tests are designed to measure reading literacy and the third and fourth to measure math literacy. Suppose the test scores have been standardized and the covariance (or in this case, the correlation) matrix of the test scores is

$$\begin{pmatrix} 1.00 & & & \\ .49 & 1.00 & & \\ .25 & .25 & 1.00 & \\ .25 & .25 & .49 & 1.00 \end{pmatrix},$$

and a possible two-factor solution is

$$\Lambda = \begin{pmatrix} .61 & .35 \\ .61 & .35 \\ .61 & -.35 \\ .61 & -.35 \end{pmatrix}, \Phi = \begin{pmatrix} 1.00 & \\ .00 & 1.00 \end{pmatrix},$$
$$= \begin{pmatrix} .51 & & & \\ & .51 & & \\ & & .51 & \\ & & & .51 \end{pmatrix}.$$

Some authors (e.g., Thissen & Wainer, 2001) have referred to this arrangement of factor loadings as the *principal axis orientation*. The first factor in this solution accounts for the most variance, and the second factor accounts for most of the additional variance that is independent of the first factor, and so forth. The factors are orthogonal (uncorrelated), as reflected by the Φ matrix. The unique variances are all equal to .51, indicating that 49% of the variances of the observed variables are associated with the two common factors.

The two-factor solution highlights the indeterminacies described earlier. First, the scale of the latent factors is indeterminate. Assigning a scale for each latent variable is one aspect of model identification. For instance, the following solution fits the observed correlations equally well:

$$\Lambda = \begin{pmatrix} .30 & .18 \\ .30 & .18 \\ .30 & -.18 \\ .30 & -.18 \end{pmatrix}, \Phi = \begin{pmatrix} 4.00 & \\ .00 & 4.00 \end{pmatrix},$$
$$= \begin{pmatrix} .51 & & & \\ & .51 & & \\ & & .51 & \\ & & & .51 \end{pmatrix}.$$

Note that an increase in the common factor variances is canceled out by a corresponding decrease in factor loadings, leaving the implied covariance matrix unaltered. It is customary to standardize the latent variables, but it is by no means a requirement. Second, for a solution with more than one factor, the factor loadings are rotationally indeterminate. That is, there exist an infinite number of loading matrices that all reproduce the observed correlation matrix equally well. Thurstone's (1947) simple structure criterion is often used to resolve the indeterminacy. For instance, the solution

$$\Lambda = \begin{pmatrix} .68 & .18 \\ .68 & .18 \\ .18 & .68 \\ .18 & .68 \end{pmatrix}, \Phi = \begin{pmatrix} 1.00 & \\ .00 & 1.00 \end{pmatrix},$$
$$= \begin{pmatrix} .51 & & & \\ & .51 & & \\ & & .51 & \\ & & & .51 \end{pmatrix}$$

represents an orthogonal simple structure arrangement. In fact, it is in Kaiser's (1958) varimax orientation. Simple structure eases the interpretation of the factors. Take the first factor, for instance. The two reading literacy tests load highly on this factor, but the math tests have low loadings. This factor could therefore be termed *reading literacy*. In a similar way, the second factor could be named *math literacy*. Thurstone's (1947) simple structure criterion is not restricted to uncorrelated factors. Indeed, the solution

$$\Lambda = \begin{pmatrix} .70 & .00 \\ .70 & .00 \\ .00 & .70 \\ .00 & .70 \end{pmatrix}, \Phi = \begin{pmatrix} 1.00 & \\ .50 & 1.00 \end{pmatrix},$$

$$= \begin{pmatrix} .51 & & & \\ & .51 & & \\ & & .51 & \\ & & & .51 \end{pmatrix}$$

yields an even cleaner interpretation. In this solution, the reading and math literacy factors are correlated at .50. Eliminated are the small but nonzero cross-loading of reading tests on the math factor and the cross-loading of math tests on the reading factor. The arrangement shown here uses Jennrich's (1966) direct quartimin rotation method.

Graphically, rotation literally means a reorientation (rotation) of the coordinate axes in the factor space. Before analytical rotation methods became available, rotation was done graphically, by hand. As Figure 5.3 shows, the observed variables are the arrows (vectors) living in the factor space (in this case, two dimensional). The dashed lines represent a pair of orthogonal coordinate axes, one for the reading factor and the other for the math factor. The vectors point from the origin (0,0) to points defined by the factor loadings. For instance, the dotted-and-dashed vectors are pointing in the principal axes orientation, whereas the solid vectors are in the varimax simple structure orientation. The varimax loadings can be obtained by rotating the vectors from the principal axes orientation counterclockwise by 45 degrees. As can be seen, the varimax-rotated vectors are much closer to the axes. If the coordinate axes are allowed to be oblique (correlated), with a correlation equal to .50 (the cosine of the angle between the two axes), then a perfect independent cluster solution can be obtained by making the axes and the vectors coincide, that is, the axes go through the solid vectors.

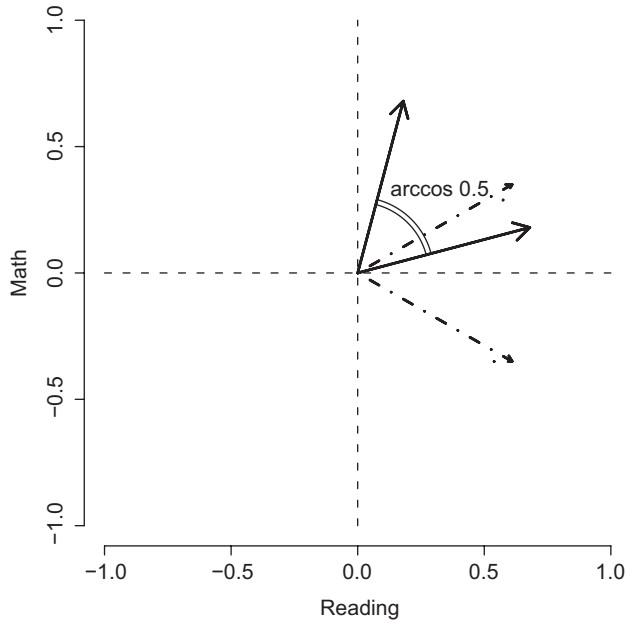

FIGURE 5.3. Factor rotation. Dashed lines represent orthogonal coordinate axes. Dotted-and-dashed vectors point in the principal axes orientation, and solid vectors are in the varimax simple structure orientation. arccos = inverse cosine function.

Estimation

After observing data from a sample of size *N*, a sample covariance matrix can be computed as $\mathbf{S} = \left(N-1\right)^{-1} \sum_{i=1}^{N} \left(\mathbf{y}_i - \bar{\mathbf{y}}\right)\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)'$, where $\bar{\mathbf{y}} = N^{-1} \sum_{i=1}^{N} \mathbf{y}_i$ is the sample mean vector based on a sample of test scores from *N* individuals. When appropriate identification conditions are met (so that the solution becomes unique), the covariance structure model can be fitted to a sample covariance matrix by minimizing the discrepancy between Σ and S. This is usually accomplished iteratively (aided by a computer) by gradually improving the estimates of factor loadings, correlations, and uniqueness from some initial values so that the resulting Σ (as a function of Λ, Φ, and Δ) becomes closest to S. Different assumptions about the distributions of the observed variables lead to different estimators. The simplest of such estimators is the ordinary least squares estimator. The ordinary least squares estimator is defined by minimizing the following discrepancy function:

$$F_{OLS}\left(\Sigma, \mathbf{S}\right) = \frac{1}{2} \mathrm{tr}\left[\left(\Sigma - \mathbf{S}\right)^2\right], \tag{5.4}$$

where tr(·) denotes the trace operator (sum of the diagonal elements). A discrepancy function measures the difference between the factor analysis model–implied covariance matrix and the observed sample covariance matrix, and it is zero if and only if Σ is the same as S. A more technical definition of

discrepancy function can be found in Browne (1984), among others. The ordinary least squares discrepancy function uses the sum of squared differences as a measure of discrepancy. This definition of discrepancy does not require assuming a specific distributional form of the observed variables.

Alternatively, an estimator with some very specific distributional form assumptions is the maximum Wishart likelihood (MWL) estimator, which is defined by minimizing the MWL discrepancy function:

$$F_{MWL}(\Sigma, \mathbf{S}) = \log|\Sigma| + \mathrm{tr}(\mathbf{S}\Sigma^{-1}) - \log|\mathbf{S}| - n, \quad (5.5)$$

where $|\cdot|$ denotes the determinant of a matrix. MWL is developed from normal theory. As such, it requires multivariate normality of the observed variables. When distributional assumptions are met, $(N - 1)$ times the minimized value of F_{MWL} is distributed asymptotically as a central chi-square variable under the null hypothesis that the factor analysis model fits exactly in the population, as shown by Bock and Bargmann (1966) and Jöreskog (1969), among many others. This is often referred to as the *overall model fit chi-square statistic*. When the model is not exactly correctly specified, the model fit chi-square statistic is distributed as a noncentral chi-square variable (Steiger, Shapiro, & Browne, 1985). The latter distributional result is particularly useful for model-fit indices that directly depend on the minimized value of F_{MWL} such as the root-mean-square error of approximation (Steiger & Lind, 1980).

MWL enjoys widespread use partly as a consequence of the success of maximum likelihood estimation in general, but blind reliance on the MWL chi-square test statistic can have adverse consequences. Setting aside issues such as non-normality, one of the most prominent problems is that the null hypothesis in the model fit test is never true (MacCallum, 2003) because all models are wrong to varying degrees (as noted by, e.g., Box, 1979). This fact implies that with a large enough sample size, the chi-square statistic is always going to lead to a rejection of the model. Cudeck and Henly (1991) discussed this sample size "problem" in detail and offered remedies. However, MWL may not be an ideal estimation method for EFA after all. MacCallum and Tucker (1991) and Briggs and MacCallum (2003) documented conditions under which MWL fails to recover major factors but ordinary least squares continues to function as expected in EFA. Another analytically tractable and flexible estimator that has seen some use in practice is the general weighted least squares estimator (see Yuan & Bentler, 2007). Without appealing to a formal justification (specification testing of the Hausman [1978] type), in applications of factor analysis using several estimation methods is often helpful to confirm that the results are not particularly sensitive to the choice of estimators.

PRACTICAL CONSIDERATIONS FOR FACTOR ANALYSIS OF TESTS

In the previous section, the fundamental factor analysis model equations were presented. Assuming continuous observed variables, several estimation methods were discussed. Issues such as rotation were illustrated. Now, it is useful to revisit some practical aspects of factor analysis of tests.

Two Modes of Uses

Different objectives for factor analysis can lead to different techniques for data analysis. The underlying statistical model remains the same, but the exploratory and confirmatory applications of factor analysis highlight different aspects of the statistical methodology.

Exploratory factor analysis. Under the first mode, an investigator uses factor analysis as an exploratory tool to identify the number of underlying factors and their nature for a battery of psychological tests. This gives rise to EFA. Applied EFA is centered around two key issues: selecting a number of common factors and factor rotation.

Several methods have been proposed to address the number-of-factors problem. The review of Fabrigar, Wegener, MacCallum, and Strahan (1999) showed that the applied factor analysis literature is dominated by the use of decision rules with the sole aim of uncovering a true number of factors. A widely adopted (not necessarily optimal or strongly endorsed) decision rule sets the number of factors to be the number of eigenvalues that exceed 1.0 for the sample correlation matrix. Another method (the

scree test) is a graphical procedure that requires the investigator to plot the series of eigenvalues of the sample correlation matrix and identify the last major discontinuity in the sequence to ascertain the number of factors. Yet a third method is parallel analysis (Horn, 1965), which is based on a comparison of eigenvalues that one expects to find from random data with those from the observed data. More sophisticated users examine the model fit tests, for example, the MWL chi-square test. Even more sophisticated users combine this with examination of fit indices such as the Tucker–Lewis (1973) index (also known as the *nonnormed fit index*) and other indices (see the review in Browne & Cudeck, 1993) such as the root-mean-square error of approximation and expected cross-validation index. For a most clear review of significance testing and model fit indices used to decide the number of factors, please refer to Bentler and Bonett (1980).

Although some of these methods may have a certain degree of theoretical basis, a more useful principle is to realize that all models are wrong to some degree and thus no one true number of factors exists. Tucker, Koopman, and Linn (1969) and MacCallum and Tucker (1991) provided clear examples in which the true number of factors is probably far too many to be accounted for in any parsimonious manner. It is therefore useful to adopt the good-enough principle (Serlin & Lapsley, 1985) and choose such a number of factors that the model fit deteriorates markedly if fewer factors are retained and does not improve dramatically if more factors are added into the model. In practice, the heuristic procedures mentioned earlier, model fit tests and indices, and examination of residuals should be taken together. Finally, the substantive interpretability of the obtained solution should be the most important criterion in justifying the choice of the number of factors.

In EFA, analytical rotation to simple structure is carried out routinely. Fabrigar et al.'s (1999) review indicated a preference for orthogonal rotation in the applications of factor analysis. However, it is rare, if ever, that batteries of tests are designed to measure completely uncorrelated constructs. Given the conceptual superiority of oblique rotation, however, one should at least attempt both orthogonal and oblique rotation. When some prior knowledge about the nature of the factors is available, Browne (2001) recommended partially specified target rotation, which can be conducted either orthogonally or obliquely. He noted that instead of relying on heuristic rules about the magnitude of the factor loadings, the significance of factor loadings can be tested statistically with the availability of standard errors for rotated loadings in software packages such as CEFA (Browne, Cudeck, Tateneni, & Mels, 2010). Regardless of the choice of rotation methods, Browne's (2001) review makes it amply clear that much human judgment and interpretation are required to achieve satisfactory factor rotation results.

Confirmatory factor analysis. Under the second mode, an investigator has already developed a hypothesis about the number and nature of factors before the data analysis and uses CFA methods as a tool to test the psychological theory from which the hypothesized factor pattern is derived. Thus, the key focus of CFA is on specifying and testing falsifiable models. One way to distinguish EFA from CFA is that for the same number of factors, an EFA model only imposes the minimum number of constraints to achieve model identification (e.g., setting the scale of the latent variables, choosing an initial axes orientation), whereas CFA models contain not only the identification constraints but also testable substantive restrictions in the form of fixing, equality, range restrictions, or complex nonlinear dependence. Coupled with the ability to estimate factor means, CFA proves to be a flexible framework not only for the purposes originally intended (i.e., psychological measurement), but in other contexts as well, for example, latent curve models (Meredith & Tisak, 1990). In CFA, the researcher must usually specify a sufficiently large number of a priori zeros in the factor loading matrix, whereas in EFA, the lack of such a priori zeros results in rotational indeterminacy that must be resolved by rotating the loadings to simple structure. Yet even without the pitfalls of incorrectly conducted rotation, applied CFA often involves model respecification and empirically driven model specification searches (MacCallum, Roznowski, & Necowitz, 1992). MacCallum et al. (1992) strongly

recommended cross-validation using independent samples.

One of the most useful features of CFA in practice is its ability to simultaneously estimate CFA models in several populations. When this is combined with the capability to impose user-defined restrictions, CFA offers a powerful methodology for studying factorial invariance. Millsap and Meredith (2007) reviewed the historical perspectives of factorial invariance. When the inclusion of multiple groups is viewed as a rudimentary form of covariate analysis, actual covariates can be added into CFA models, further expanding their utility (see, e.g., Muthén, 1989). A prime example is the multiple indicators, multiple causes (MIMIC) model (see, e.g., Hauser & Goldberger, 1971).

Exploratory factor analysis or confirmatory factor analysis. Before Jöreskog (1969) developed the first successful and widely implemented computational algorithm for CFA, the applications of factor analysis relied predominantly on EFA methods, with Bock and Bargmann's (1966) analysis as a notable exception. However, the availability of more flexible factor rotation methods that permit the direct specification of loading patterns (e.g., partially specified target rotation; see Browne, 2001) has blurred the boundaries of exploratory and confirmatory analyses. In practice, well-executed factor analysis often involves a combination of EFA and CFA methods (e.g., Hawkley et al., 2005), with CFA conducted on one or more separate samples serving as a form of cross-validation (Browne & Cudeck, 1993). Finally, what should not be confused with either EFA or CFA is principal component analysis, a statistical procedure serving an entirely different purpose (data condensation), although it bears some algebraic similarity to EFA.

Sample Size

In the applications of factor analysis, two kinds of questions about sample size are particularly relevant. The first kind is about adequate sample size to achieve stable estimation. Many factor analysis procedures are derived under large sample conditions, and iterative estimation methods such as MWL tend to perform better when N is large. MacCallum, Widaman, Zhang, and Hong (1999) showed that there is no unequivocal rule of thumb about N. The sample size required for stable estimation actually depends on the communality of the observed variables. The second kind of questions about sample size involves statistical power computations: What is the minimum N for a particular statistical test (e.g., to distinguish a three-factor solution from a four-factor solution) so that the power of the test is equal to some established cutoff, say, .80. MacCallum, Browne, and Cai (2006); MacCallum, Browne, and Sugawara (1996); and Satorra and Saris (1985) provided methods to address such questions.

Nonnormality

Many of the inferential procedures associated with factor analysis require normality for accurate Type I error rates of the model fit test statistics or appropriate confidence interval coverage on the basis of estimated standard errors. When the normality of observed variables is suspect, it is advisable to adopt statistical corrections. Browne's (1984) asymptotically distribution-free estimator and Satorra and Bentler's (1994) scaled chi-square correction are the two most widely used alternatives under nonnormality. They provide nonnormality corrected test statistics, standard errors, and fit indices. If, however, the observed variables are not normally distributed because they are discrete item-level responses, the item factor analysis methods discussed in the next section are more appropriate.

ITEM FACTOR ANALYSIS

Thus far, the discussion has focused on factor analysis of batteries of tests. The observed variables in the statistical analyses are assumed to be continuous, often normally distributed (if MWL estimation is used). However, if one were to factor analyze categorical item-level data, item factor analysis models and methods would be required.

Item Factor Analysis Model

The normal ogive item factor analysis model presented here is based on Thurstone's (1947) common factor model and has roots in Thurstone's (1925) earlier work. For the ith person's response to the

jth item, a p-factor model is assumed for the underlying response process[1] variate y_{ij}^* such that $y_{ij}^* = \sum_{k=1}^{p} \lambda_{jk}\eta_{ik} + \varepsilon_{ij}$, where the ηs continue to denote the normally distributed latent common factors with mean zero and unit variance, λ_{jk} is the factor loading, and ε_{ij} is normally distributed with mean zero and unique variance $\sigma_j^2 = 1 - \sum_{k=1}^{p} \lambda_{jk}^2$ so that y_{ij}^* is has unit variance. The common factors and unique factors are uncorrelated.

For brevity, only the dichotomous version of the model is presented. The observed $0 - 1$ response y_{ij} is related to y_{ij}^* via a threshold parameter τ_j, also referred to as *standardized difficulty* in the education literature, such that $y_{ij} = 1$ is observed if $y_{ij}^* > \tau_j$ and $y_{ij} = 0$ otherwise. It follows that person j endorses item i (or obtains a correct response, in an educational testing context) with probability

$$P\left(y_{ij} = 1 \mid \eta_i\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z(\eta_i)} \exp\left(-\frac{t^2}{2}\right) dt, \quad (5.6)$$

$$\text{where } z\left(\eta_i\right) = \frac{\sum_{k=1}^{p} \lambda_{jk}\eta_{ik} - \tau_j}{\sigma_j}.$$

In terms of the item parameters, Bock and Aitkin (1981) used the following parameterization:

$$z(\eta_i) = \alpha_j + \sum_{k=1}^{p} \beta_{jk}\,\eta_{ik},$$

where $\alpha_j = -\tau_j / \sigma_j$ is the item intercept, and $\beta_{jk} = \lambda_{jk} / \sigma_j$ is called an *item slope*. The αs and βs are also known as the *unstandardized parameters*, whereas the τs and λs are the *standardized parameters*. In practice, maximum likelihood estimation of the item factor analysis model often involves a logistic substitution. That is, instead of Equation 5.6, the probability of endorsement or a correct response is

$$P\left(y_{ij} = 1 \mid \eta_i\right) = \frac{1}{1 + \exp\left[-D\left(\alpha_j + \sum_{k=1}^{p} \beta_{jk}\eta_{ik}\right)\right]}, \quad (5.7)$$

where D is a scaling constant (often equal to 1.702) such that the logistic cumulative distribution function becomes virtually identical in shape to the normal cumulative distribution function in Equation 5.6.

Note that this item factor analysis model is still rotationally indeterminate unless the analysis is done in a confirmatory mode, with a sufficient number of fixed zero loadings. More general versions of the model for ordinal and nominal data are straightforward extensions (see, e.g., Cai, 2010a, 2010b; Thissen, Cai, & Bock, 2010). Multiple group versions of these models are also available (see, e.g., Bock & Zimowski, 1997; Cai et al., 2011).

Estimation

Estimating the parameters of the item factor analysis model is a nontrivial task. The researcher must deal with a multiway contingency table formed by the cross-tabulations of the item responses and, at the same time, tackle the numerical integration problem that results in an exponentially increasing computational burden in the number of factors for full-information methods based on raw item response data and (unless limited-information shortcuts are taken) an exponential increase in the number of items for methods based on tetrachoric or polychoric correlations. As reviewed by Wirth and Edwards (2007), there are two basic classes of estimation methods: those that identify the parameters from lower order marginal tables (limited information) and those that identify the parameters directly from the raw data (full information).[2]

Limited-information methods. Typically, limited-information methods refer to a multistage estimation procedure in which the categorical item responses are first used to estimate a tetrachoric–polychoric correlation matrix (depending on item type), along with the full asymptotic covariance matrix of the tetrachoric–polychoric correlations (e.g. Muthén, 1978) or only a part of it, such as the diagonal elements. In some cases, thresholds are estimated separately even before the first stage by inverting the observed category proportions using the inverse normal cumulative distribution function. In the second stage, the correlations are analyzed in a standard factor analysis software program by the weighted least squares method with the inverse of the asymptotic covariance matrix serving as the weights. There are several variations on this basic

[1]The word *process* in *response process* is somewhat archaic English meaning *number*.

[2]Testlet analysis based on the nominal categories model (Thissen, Steinberg, & Mooney, 1989) also uses full-information estimation.

setup, but they share the feature that the estimation of the interitem tetrachoric–polychoric correlation matrix is accomplished in a pairwise manner. Hence, each tetrachoric–polychoric correlation only draws information from a bivariate subtable of the item response cross-classifications. Although this may be computationally advantageous, replacing the full *n*-dimensional integral with a set of bivariate integrations, the tetrachoric–polychoric correlation matrix thus obtained may not be positive definite. Methods such as Fraser and McDonald's (1988) NOHARM also use limited information but do not involve the computation of correlations. Wirth and Edwards (2007) noted that limited-information methods tend to be more useful when the number of items is small.

Full-information methods. Pioneered by Bock, Gibbons, and Muraki (1988), current full-information methods use either maximum likelihood or Bayesian estimation to obtain loading and threshold estimates directly from raw data. This line of work is principally developed out of multidimensional IRT (see, e.g., Reckase, 2009). Full-information methods can handle a more flexible class of item responses but are much more computationally demanding.

General Hierarchical Item Factor Models

A special class of item factor models deserves separate comments. This class includes Gibbons and Hedeker's (1992) bifactor model (see also Gibbons et al., 2007); Wainer, Bradlow, and Wang's (2007) testlet response theory model; Cai's (2010c) two-tier item factor model; and Cai et al.'s (2011) generalized bifactor model, just to name a few. These models tend to have a hierarchically arranged factor pattern, with one or more primary dimensions that can influence all items and a set of mutually orthogonal specific dimensions that only influence specific and nonoverlapping subsets of items, accounting for residual dependence after the extraction of the primary dimensions. Frequently, the hierarchical factor pattern not only provides a clean interpretation of the construct or constructs that the items are purported to measure but, more important, leads to highly efficient full-information estimation methods that can dramatically reduce the dimensionality of the model.

Practical Considerations

Because of the relatively more technical nature of the full-information approach, two practical issues must be considered. The first involves parameter estimation algorithms, which tend to be computationally intensive. The second is model fit evaluation, which remains an area of active research.

Algorithms for full-information estimation. To the practitioner, full information methods contain a bewildering array of estimation algorithms. For a small number of factors, Bock and Aitkin's (1981) expectation–maximization (EM) algorithm tends to be a robust choice. The generalized dimension reduced EM algorithm (Cai, 2010c) is efficient for hierarchical models. For a small to medium number of factors, Schilling and Bock's (2005) adaptive quadrature-based EM algorithm can be substantially more efficient than Bock and Aitkin's (1981) EM algorithm using fixed quadrature. Computation of standard errors in item factor analysis has been a chronic problem because of the reliance on EM algorithms that do not provide standard errors on convergence. To address this, Cai (2008) offered a method to compute standard errors for IRT models and item factor analysis models using a supplemented EM algorithm.[3]

Researchers have recently devoted increased attention to the development of more efficient and flexible algorithms for high-dimensional item factor analysis, from both Bayesian and likelihood approaches. Either Bayesian Markov chain Monte Carlo methods must be used (e.g., Edwards, 2010) or a Monte Carlo–based optimization algorithm such as the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010a, 2010b) should be used for maximum likelihood estimation.

Model evaluation. In comparison with linear factor analysis of tests, model evaluation and model fit assessment for item factor analysis are in a much less developed state. The main problem is the inherent sparseness of the underlying multiway contingency table. In practice, overall goodness-of-fit tests that are based on limited information (e.g., Cai, Maydeu-Olivares, Coffman, & Thissen,

[3]The supplemented EM algorithm is implemented in IRTPRO (Cai, Thissen, & du Toit, 2011).

2006; Maydeu-Olivares & Joe, 2005) seem to be much more promising than the traditional Pearson chi-square or likelihood ratio chi-square statistics. Under appropriate conditions, the likelihood ratio chi-square difference test can still be used to gauge the relative fit of nested models (Maydeu-Olivares & Cai, 2006). Among tests that can diagnose sources of model misfit, the item fit tests developed by Orlando and Thissen (2000) and the local dependence indices developed by W. H. Chen and Thissen (1997) seem more promising.

CONCLUDING COMMENTS

Factor analysis seeks to improve the understanding of the structure of psychological tests. Factor analysis can be applied to batteries of tests, or it can be applied to test items directly. Tracing the historical developments from Spearman's (1904) initial contributions, an overview of both linear common factor analysis and nonlinear item factor analysis was provided.

From its inception, factor analysis has been a method that is studied, applied, and scrutinized by both psychologists and statisticians. After more than 100 years, it still represents a lively area in the statistical methodology literature. When applied to sets of psychological tests, it has deepened psychologists' understanding of psychological tests and assessments. When applied to a set of test items, it clarifies the underlying dimensionality of the test and can aid the construction of better measurement instruments. Even more broadly, factor analysis paved the way for the success of general latent variable modeling frameworks such as structural equation modeling. As Cudeck and MacCallum (2007) noted, factor analysis is indeed a success story in both psychology and statistics.

References

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London, England: Arnold.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606. doi:10.1037/0033-2909.88.3.588

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459. doi:10.1007/BF02293801

Bock, R. D., & Bargmann, R. (1966). Analysis of covariance structures. *Psychometrika, 31*, 507–534. doi:10.1007/BF02289521

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261–280. doi:10.1177/014662168801200305

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer-Verlag.

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association, 74*, 1–4.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38*, 25–56. doi:10.1207/S15327906MBR3801_2

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83. doi:10.1111/j.2044-8317.1984.tb00789.x

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*, 111–150. doi:10.1207/S15327906MBR3601_05

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York, NY: Plenum.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & S. Long (Eds.), *Testing structural equation models* (pp. 131–161). Newbury Park, CA: Sage.

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2010). CEFA: Comprehensive exploratory factor analysis (Version 3.04) [Computer software]. Retrieved from http://faculty.psy.ohio-state.edu/browne/software.php

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309–329. doi:10.1348/000711007X249603

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75,* 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis.

Journal of Educational and Behavioral Statistics, 35, 307–335. doi:10.3102/1076998609353115

Cai, L. (2010c). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581–612. doi:10.1007/s11336-010-9178-0

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173–194. doi:10.1348/000711005X66419

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221–248.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189–225. doi:10.1207/s15327906mbr4102_5

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109,* 512–519. doi:10.1037/0033-2909.109.3.512

Cudeck, R., & MacCallum, R. C. (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 62*, 355–366. doi:10.1111/1467-9868.00236

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika, 75*, 474–497.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299. doi:10.1037/1082-989X.4.3.272

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6*, 323–329. doi:10.1007/BF02288588

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267–269. doi:10.1207/s15327906mbr2302_9

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . ., Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4–19. doi:10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika, 57*, 423–436. doi:10.1007/BF02295430

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Mahwah, NJ: Erlbaum.

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika, 6*, 67–77. doi:10.1007/BF02292175

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: University of Chicago Press.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. In L. Costner (Ed.), *Sociological methodology* (pp. 81–177). San Francisco, CA: Jossey-Bass.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*, 1251–1271. doi:10.2307/1913827

Hawkley, L. C., Browne, M. W., & Cacioppo, J. T. (2005). How can I connect with thee? Let me count the ways. *Psychological Science, 16*, 798–804. doi:10.1111/j.1467-9280.2005.01617.x

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41–54. doi:10.1007/BF02287965

Horn, J. L. (1965). A rationale and technique for estimating the number of factors in factor analysis. *Psychometrika, 30*, 179–185. doi:10.1007/BF02289447

Jennrich, R. I. (1966). Rotation for simple loadings. *Psychometrika, 31*, 313–323. doi:10.1007/BF02289465

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187–200. doi:10.1007/BF02289233

Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London, England: Butterworth.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). New York, NY: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38,* 113–139. doi:10.1207/S15327906MBR3801_5

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods, 11,* 19–35. doi:10.1037/1082-989X.11.1.19

MacCallum, R. C., Browne, M. W., & Sugawara, H. H. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149. doi:10.1037/1082-989X.1.2.130

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504. doi:10.1037/0033-2909.111.3.490

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502–511. doi:10.1037/0033-2909.109.3.502

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99. doi:10.1037/1082-989X.4.1.84

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*, 55–64. doi:10.1207/s15327906mbr4101_4

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009–1020. doi:10.1198/016214504000002069

McDonald, R. P. (1985). *Factor analysis and related methods.* Mahwah, NJ: Erlbaum.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107–122. doi:10.1007/BF02294746

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 131–152). Mahwah, NJ: Erlbaum.

Muthén, B. (1978). Contributions of factor analysis to dichotomous variables. *Psychometrika, 43*, 551–560.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64. doi:10.1177/01466216000241003

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. doi:10.1007/978-0-387-89976-3

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83–90. doi:10.1007/BF02294150

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533–555.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53–61. doi:10.1007/BF02289209

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*, 73–83. doi:10.1037/0003-066X.40.1.73

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201–293. doi:10.2307/1412107

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50*, 253–263. doi:10.1007/BF02294104

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 43–75). New York, NY: Taylor & Francis.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247–260. doi:10.1111/j.1745-3984.1989.tb00331.x

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Erlbaum.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451. doi:10.1037/h0073357

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34*, 421–459. doi:10.1007/BF02290601

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10. doi:10.1007/BF02291170

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511618765

Wherry, R. J. (1959). Hierarchical factor solutions without rotation. *Psychometrika, 24*, 45–51. doi:10.1007/BF02289762

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79. doi:10.1037/1082-989X.12.1.58

Yuan, K.-H., & Bentler, P. M. (2007). Structural equation modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26. Psychometrics* (pp. 297–358). Amsterdam, the Netherlands: North-Holland.

Yung, Y. F., McLeod, L. D., & Thissen, D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113–128. doi:10.1007/BF02294531

Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129–152. doi:10.1007/BF02294532

CHAPTER 6

APPLYING UNIDIMENSIONAL ITEM RESPONSE THEORY MODELS TO PSYCHOLOGICAL DATA

Steven P. Reise, Tyler M. Moore, and Mark G. Haviland

The application of unidimensional item response theory (IRT; de Ayala, 2009; Embretson & Reise, 2000) measurement models is generally standard practice in cognitive aptitude assessment. Moreover, with the developments in psychometric theory and the increase in computing power over the past 25 years, researchers are now able to apply complex statistical models to ordinal item response data in ways not imagined in the past; as such, IRT applications in noncognitive assessment are increasingly common and effective (Cella et al., 2007; Reise & Waller, 2009).

IRT modeling offers researchers many clear advantages over traditional psychometric practices (see Morizot, Ainsworth, & Reise, 2007; Reise, Ainsworth, & Haviland, 2005; Reise & Henson, 2003). Noteworthy examples are (a) a more extensive evaluation of a measure's psychometric properties, (b) a better metric for scaling individual differences and comparing group means, (c) a foundation for evaluating differential item and test functioning (i.e., bias), and (d) a basis for more informed scale and item bank development as well as short forms and computerized adaptive tests. Although these advantages are attractive, researchers wishing to apply IRT models to noncognitive item response data face a number of obstacles (Reise, 2010; Reise & Moore, 2012). Fitting an IRT model, for example, requires extensive technical knowledge. Moreover, item response data need to be consistent with several difficult-to-meet assumptions. In contrast, traditional psychometric methods are relatively simple to understand, and the assumptions are few and easy to meet.

This chapter is not a complete introductory summary of item response modeling and its virtues and potential applications. This information is available in the articles cited earlier. Instead, we provide a tutorial for empirically exploring the degree to which item response data are consistent with a unidimensional IRT model. Our guide is divided into two sections: In the first, we review commonly encountered unidimensional IRT models for dichotomous and polytomous item responses, and in the second, we present the assumptions of IRT modeling and demonstrate the process of fitting IRT models to noncognitive item response data; notably, empirically evaluating whether the data are appropriate for such modeling. In this evaluation, researchers must consider the nature of the construct being assessed as well as information about monotonicity, dimensionality, and local dependence.

ITEM RESPONSE THEORY MODELS

IRT models can be grouped according to several different characteristics, and one important way is whether they are designed for dichotomous or

This work was supported by the Consortium for Neuropsychiatric Phenomics: National Institutes of Health Roadmap for Medical Research Grants UL1-DE019580 (principal investigator, Robert Bilder) and RL1DA024853 (principal investigator, Edythe London), the National Institutes of Health Roadmap for Medical Research Grant AR052177 (principal investigator, David Cella), National Cancer Institute Grant 4R44CA137841-03 (principal investigator, Patrick Mair) for item response theory software development for health outcomes and behavioral cancer research, and the Institute of Educational Sciences Grant 00362556 (project director, Noreen Webb). The content is solely the responsibility of the authors, however, and does not necessarily represent the official views of the funding agencies.

DOI: 10.1037/14047-006
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

polytomous items. *Dichotomous items* have only two response options (e.g., correct vs. incorrect, yes vs. no, agree vs. disagree), which are common in cognitive testing where items are often graded as simply correct or incorrect. *Polytomous items* have more than two response options, as in the case of a rating scale with, for example, five agree/disagree options (*strongly agree, agree, neither agree nor disagree, disagree*, and *strongly disagree*). In the sections that follow, we begin with dichotomous IRT models and then describe how these models are changed and expanded to handle polytomous items.

Item Response Theory Models for Dichotomous Item Responses

The fundamental unit of a parametric IRT model is a mathematical equation to accurately capture the relationship between a continuous latent variable measured by a scale and the probability of endorsing an item (responding "correct" or in the keyed direction). Such functions are referred to as *item response curves* (IRCs), and variations on IRT models are nothing more than different equations to describe an IRC. Today, most software programs estimate the parameters of so-called "logistic" models; for example, Equation 6.1 is the unidimensional (one latent trait) three-parameter logistic model (3PL).

$$P(x = 1 \mid \theta) = c + (1 - c)\frac{\exp\left[a(\theta - b)\right]}{1 + \exp\left[a(\theta - b)\right]}. \qquad (6.1)$$

In Equation 6.1, θ represents individual differences on a latent variable, which can in turn be identified by assuming its mean is 0 and variance is 1 in the population. The *a* (or "slope") parameter determines the steepness of the IRC (see Figure 6.1). Items with higher *a* are more discriminating or psychometrically informative. The *b* parameter is an item location parameter, which typically ranges between −2.5 and 2.5. Items with high endorsement rates have negative location parameters and IRCs shifted to the left. Items with low endorsement rates have positive location parameters and IRCs shifted to the right (see Figure 6.1). Finally, *c* determines the lower asymptote of the IRC. The *c* parameter is primarily used to model multiple-choice aptitude test items on which even individuals who are low on the latent variable can obtain a correct answer by

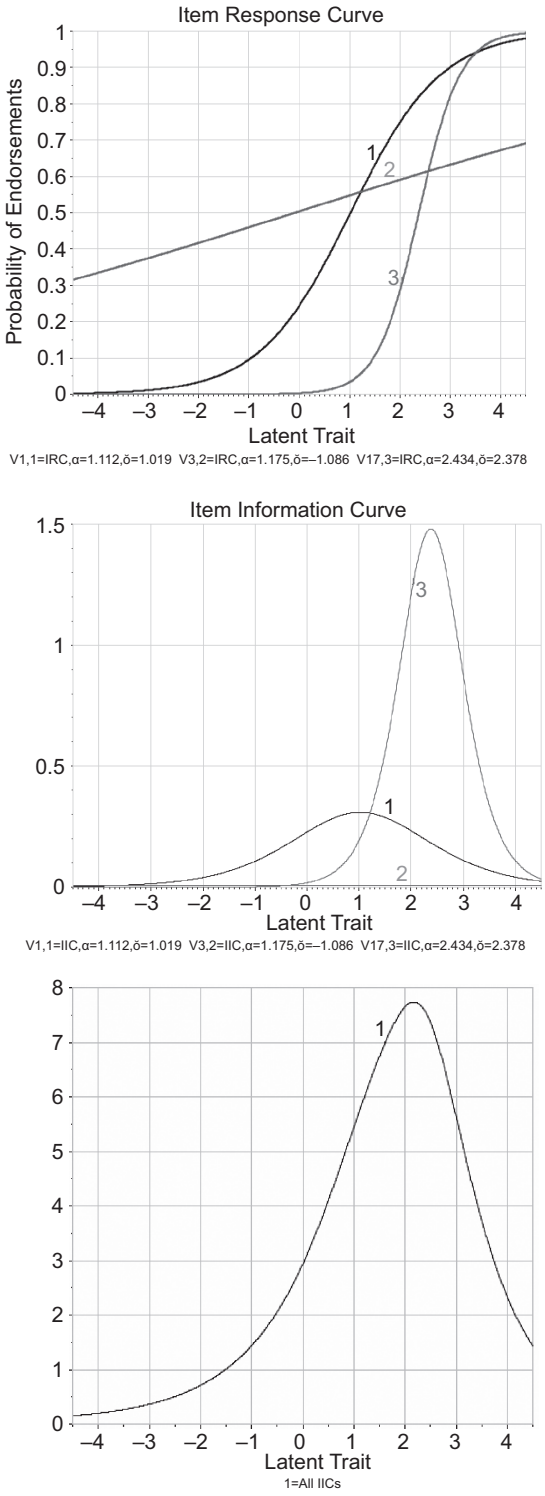

FIGURE 6.1. Item response curves (top panel) and item information curves (middle panel) for three BIS items (3, 4, and 17) and the scale information curve (bottom panel) for the entire BIS measure.

guessing or some other means than actually knowing the answer.

For items taken from noncognitive measures, on which individuals are not expected to guess at the keyed answer, a more restricted two-parameter logistic (2PL) model may be appropriate to describe the IRC. In this model, there is no lower asymptote parameter, and, thus, the probability of endorsing the item goes toward zero for people in the lowest trait ranges. The 2PL model is shown in Equation 6.2.

$$P(x = 1 \mid \theta) = \frac{\exp\left[a(\theta - b)\right]}{1 + \exp\left[a(\theta - b)\right]}. \qquad (6.2)$$

In this model, items are allowed to vary in two ways—slope (a) and location (b). Equation 6.2 states that the conditional probability of endorsing the item is a weighted (by the a parameter) function of the difference between an individual's standing on the latent variable (θ) and the item's location parameter (b). As in Equation 6.1, a reflects the discrimination of the item—higher values reflect more discriminating items. Moreover, b determines the location of the IRC and indicates the point on the latent variable continuum at which the probability of endorsing an item is .50. Thus, if b is 0.75, then individuals with latent trait values below 0.75 will have a less than 50/50 chance of endorsing the item, whereas individuals with latent trait values above 0.75 will have a greater than 50/50 chance of endorsing the item.

Finally, if one is willing to assume that all the items in a measure are equally discriminating, then the item slope parameters can be constrained to be a constant across items. Such a one-parameter logistic (1PL) model is shown in Equation 6.3.

$$P(x = 1 \mid \theta) = \frac{\exp\left[\ddot{a}(\theta - b)\right]}{1 + \exp\left[\ddot{a}(\theta - b)\right]}. \qquad (6.3)$$

In this model, $\ddot{a}$ means that the slopes are constrained to equality across items; in other words, the items are not allowed to vary in their discrimination. The b location parameter is interpreted exactly as before, that is, as the point on the latent variable scale at which the probability of endorsement is .50.

To illustrate the application of these models, we used a sample of 401 community members who responded to the 30-item Barratt Impulsivity Scale (BIS) Version 11 (Barratt, 1959; Patton, Stanford, & Barratt, 1995). Item responses were dichotomized for the sake of these analyses. Coefficient alpha was .76, with an average item intercorrelation of .10. Listed in Table 6.1 are the item descriptive statistics (item means, item–test correlations) and estimated IRT item parameters from Equation 6.2 (2PL). All analyses in Table 6.1 and subsequent figures were done with mvIRT (Multivariate Software, Inc., 2010), and all program defaults (full information or marginal maximum likelihood estimation methods) were used.

TABLE 6.1

Classical Descriptives and Item Response Theory Parameter Estimates for the Barratt Impulsivity Scale Version 11

	Descriptives		IRT model (2PL)	
Item	*M*	Item–test *r*	*a*	*b*
1	0.31	.48	1.11	0.92
2	0.12	.42	1.69	1.68
3	0.52	.20	0.18	–0.49
4	0.40	.19	0.12	3.51
5	0.08	.25	0.91	3.08
6	0.18	.33	0.94	1.88
7	0.37	.42	0.87	0.72
8	0.16	.47	1.46	1.52
9	0.29	.55	1.54	0.81
10	0.52	.39	0.68	–0.13
11	0.16	.33	0.88	2.18
12	0.17	.48	1.54	1.40
13	0.49	.44	0.93	0.08
14	0.13	.38	1.34	1.82
15	0.43	.28	0.34	0.90
16	0.28	.30	0.57	1.78
17	0.21	.54	2.43	0.98
18	0.14	.36	1.07	2.02
19	0.25	.44	1.49	1.01
20	0.23	.50	1.44	1.14
21	0.18	.25	0.55	2.87
22	0.14	.34	0.98	2.14
23	0.14	.22	0.38	4.84
24	0.09	.30	0.97	2.71
25	0.16	.38	1.04	1.91
26	0.25	.40	1.10	1.23
27	0.39	.22	0.29	1.58
28	0.17	.36	1.09	1.79
29	0.52	.22	0.19	–0.39
30	0.37	.43	0.77	0.79

Note. IRT = item response theory; 2PL = two-parameter logistic.

Several noteworthy findings are shown in Table 6.1. First, as with any IRT application, there are links between classical test theory indices and IRT parameter estimates. For example, the item means correspond to the item locations in the 2PL (low means → high locations), and the item–test correlations correspond to the item slopes (high correlations → high slopes). Second, in the 2PL, the low item slopes for Items 3 and 4 mean that these items contribute little to the measurement of the common latent variable (ostensibly, impulsivity), whereas the relatively high slope for Item 17 shows that this is a critical or highly differentiating item.

For illustration, in the top panel of Figure 6.1, we show the IRCs for three items that vary in slope and location. The slope and location parameter estimates are essential in IRT modeling because they determine the amount and location of psychometric information. *Item information* refers to how well responses to an item differentiate or discriminate among individuals along the latent variable continuum; items with a high slope provide more information, and where that information is located is determined by the location parameter. In the middle panel of Figure 6.1, for example, we show the item information curves for the three items in the top of Figure 6.1. Clearly, Item 17 provides more information in the high range of the latent variable, whereas Item 3 provides essentially no information.

Finally, if one assumes our data meet IRT modeling assumptions, then item information curves are additive across items within a measure. (Note that, in the next section, we provide details on the various IRT assumptions and the ways in which they can be evaluated.) Researchers can thus calculate a scale information curve that illustrates how informative an item set is across the latent variable continuum. In the bottom of Figure 6.1, we show the scale information curve for the BIS measure. Clearly, this function is peaked at the high end, showing that the measure provides its best precision in the high (impulsivity) trait range. The scale has a relatively limited amount of information in the low trait range; in other words, the measure does not differentiate among low-trait individuals. The peaked information in the high range is not surprising given that the measure was designed to differentiate impulsive from nonimpulsive individuals rather than to scale people precisely from impulsive to high constraint.

The importance of scale information is best understood by recognizing that an individual's standard error of measurement is an inverse function of the information. Specifically, a conditional standard error is one divided by the square root of the information. In this scale, the maximum information is around 8, so the standard error of measurement for individuals in that trait range is $1 / \sqrt{8} = .356$. Note that the standard errors are even larger for individuals scoring in trait ranges away from the peak of the scale information function. The overall low scale information values across the latent variable range are not surprising given that even with 30 items, coefficient alpha is only .76. Stated in different terms, the items are not highly intercorrelated (average $r = .10$), suggesting that what they share in common (impulsivity) is only weakly measured by these items.

Item Response Theory Models for Polytomous Item Responses

Noncognitive assessment (e.g., personality and health outcomes) relies heavily on questionnaires with multipoint or polytomous item response formats. Such data require slightly more sophisticated IRT models. There are numerous polytomous IRT models (see Nering & Ostini, 2010; Ostini & Nering, 2006), but in this chapter, we describe only the graded response model (GRM; Samejima, 1969) because of its popularity in personality, psychopathology, and health outcomes research (Reise & Waller, 2009).

To illustrate the GRM, we use an eight-item set ($N = 1{,}168$) of polytomous responses from an Extraversion parcel (E1: Warmth) from the revised NEO Personality Inventory (NEO PI–R; Costa & McCrae, 1992). Coefficient alpha in this college student sample was .76 with an average interitem correlation of .28. A subsample from this dataset has previously been analyzed in Reise and Henson's (2000) simulation of computerized adaptive testing with the NEO PI–R. In these examples, our emphasis is on the item parameter estimates and evaluations of model assumptions. Moreover, because of the very small number of responses in the lowest category for some items, we collapsed the first and

second response categories (0 and 1), resulting in four category responses (0–1–2–3) instead of the original five (0–1–2–3–4). The new response labels were 0 = *strongly disagree* and *disagree* combined, 1 = *neutral,* 2 = *agree,* and 3 = *strongly agree.*

Graded Response Model

The chief objective of polytomous IRT models is to estimate a set of best-fitting category response curves (CRCs). These CRCs represent the relationship between a person's latent trait level (θ) and the probability of responding in a particular category. It is easy to intuit the GRM in that it is simply a generalization of the 2PL (Equation 6.2) to account for the multiple thresholds between response categories in a polytomous item. For a dichotomous item, there is only one threshold between the response options, and thus only a single equation is needed to describe the probability of responding above that threshold (i.e., transitioning from a 0 to a 1 response), and only a single location parameter needs to be estimated. For a multipoint item, however, there are the number of categories (k) minus 1 thresholds ($m = k - 1$) between the response options. One thus needs to estimate m 2PL models with the constraint that the slope parameters are equivalent within items. These 2PL models, one for each dichotomy (e.g., for a 4-point item: 0, vs. 1, 2, 3; 0, 1, vs. 2, 3; 0, 1, 2, vs. 3), are called *threshold response curves* (TRCs). The TRCs represent the probability of responding above a between-category threshold as a function of the latent variable. These functions, however, are not the same as the functions describing CRCs. CRCs must be derived indirectly from the TRCs, which is why the GRM is classed as an indirect IRT model (Embretson & Reise, 2000, pp. 96–98). The indirect, two-step process required to derive the CRCs of the GRM is as follows.

The first step is to estimate a set of m TRCs for each item, where k is the number of response categories. These TRCs are represented by Equation 6.4, which is identical to Equation 6.2 but with equal slopes within each item.

$$P_x^*(\theta) = \frac{\exp\left[a\left(\theta - b_j\right)\right]}{1 + \exp\left[a\left(\theta - b_j\right)\right]}, \tag{6.4}$$

where $j = 1 \ldots m$. By definition, the probability of responding in or above the lowest response category is $P_{(x=0)}^*(\theta) = 1.0$, and the probability of responding above the highest response category is $P_{(x=4)}^*(\theta) = 0.0$. The second step involves using the information provided by Equation 6.4 (TRCs) to describe the relationships (CRCs) between each response category and the latent variable, and the CRCs can be derived by simple subtraction of TRCs:

$$P_x(\theta) = P_{(x)}^*(\theta) - P_{(x+1)}^*(\theta). \tag{6.5}$$

To illustrate the GRM, Table 6.2 displays the estimated item parameters for the set of NEO–PI–R items scored with four response options.

TABLE 6.2

Response Frequencies, Item Statistics, and Graded Response Model Item Parameter Estimates for the Revised NEO Personality Inventory E1 (Warmth) Data

	Response frequencies						IRT graded response model			
Item	0	1	2	3	*M*	Item–test *r*	*a*	*b1*	*b2*	*b3*
1	0.11	0.22	0.53	0.14	1.7	.58	1.12	–2.23	–0.78	1.96
2	0.05	0.08	0.46	0.41	2.2	.63	1.61	–2.44	–1.65	0.30
3	0.04	0.13	0.54	0.29	2.1	.65	1.48	–2.70	–1.44	0.85
4	0.08	0.14	0.42	0.35	2.1	.65	1.41	–2.24	–1.21	0.58
5	0.05	0.11	0.49	0.35	2.1	.71	2.14	–2.14	–1.31	0.48
6	0.17	0.19	0.41	0.23	1.7	.59	1.21	–1.63	–0.60	1.28
7	0.07	0.12	0.48	0.34	2.1	.53	0.91	–3.28	–1.91	0.86
8	0.08	0.20	0.62	0.11	1.8	.54	0.95	–2.97	–1.20	2.56

Note. IRT = item response theory.

The program IRTPRO (Cai, Thissen, & du Toit, 2011) was also used to estimate GRM parameters, and the figures shown are graphs taken directly from IRTPRO output. Note that as with dichotomous IRT models, there is a relationship between the IRT parameter estimates and the traditional statistics in polytomous models. For example, Item 5 has the highest item–test correlation and IRT slope (a = 2.14), and Item 7 has the lowest item–test correlation and IRT slope (a = 0.91). Although there is not much variation in item means, it is clear from Table 6.2 that category response frequencies correspond roughly to the values of the threshold parameter estimates. Generally speaking, polytomous items with relatively low means have thresholds shifted to the right, whereas relatively higher means have thresholds shifted to the left.

Figure 6.2 shows the CRCs for two NEO PI–R items, Item 5 with relatively high discrimination and Item 8 with relatively low discrimination. The first thing that is apparent in these figures is that the CRCs mimic the response frequencies fairly well.

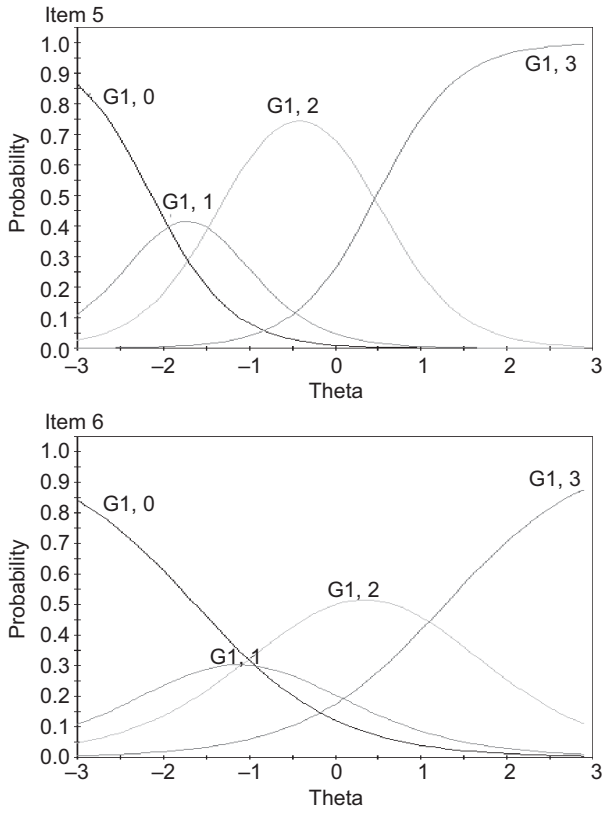

FIGURE 6.2. Category response curves for two E1 items (5 and 8) that vary in discrimination.

That is, for Item 5, most people responded in Category 2, and thus the CRC for this category covers a broad range of the latent variable. Also notice that the CRCs are shifted to the left for Item 8 relative to Item 5 (because it is an "easier" item) and that the CRCs are more sharply distinguished for Item 5 relative to Item 8 (because it is a more discriminating item).

In Figure 6.3, we show the item information for these two items. Generally, because of the multipoint response format, polytomous items tend to provide more item information, and that item information tends to be better spread out across the latent variable continuum. In these data, as expected, Item 5 provided substantially more information than did Item 8. Finally, the scale information is shown in the bottom panel of Figure 6.3. As with the BIS (dichotomous example), the scale information for this NEO PI–R measure is rather low for all trait ranges, and thus the standard errors of measurement are high for all trait ranges. The standard errors tend to range between .4 and .5 for most individuals and then increase for individuals with high trait-level estimates. Owing to the polytomous response format, the scale information tends to spread out somewhat, although it is still clearly peaked.

APPLYING ITEM RESPONSE THEORY MODELS

There are many good reasons to apply an appropriate IRT model to a particular data set, but before doing so one should always confirm that certain criteria are met. Almost all such criteria have to do with characteristics of the data itself, such as whether the assumptions of IRT are judged to have been met. In the sections that follow, we list and describe the assumptions of IRT and explain why they are important.

Are the Data Appropriate for Item Response Theory?

Most researchers are likely familiar with the basics of traditional scale development and evaluation (e.g., true scores, standard errors of measurement, item–test correlations, coefficient alpha internal consistency estimates). As such, these basic principles and their respective conventional standards

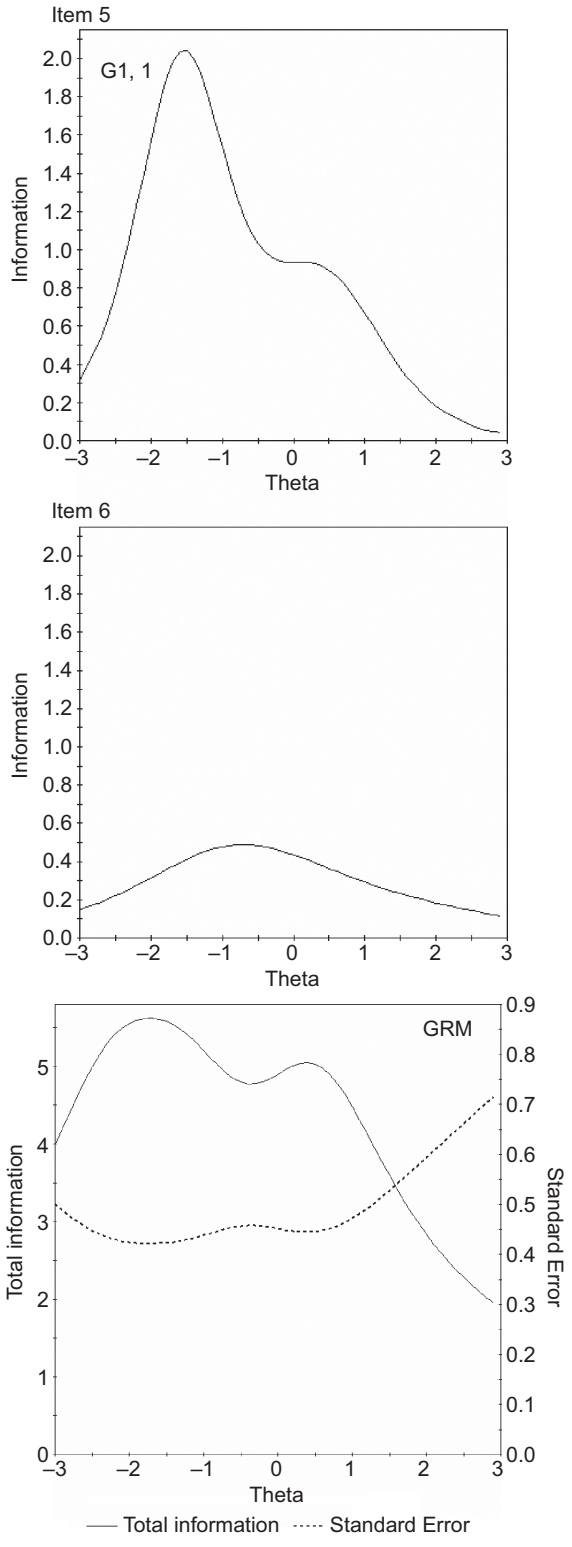

FIGURE 6.3. Item information curves for two E1 items (5 and 8) that vary in discrimination (top panel) and scale information curve for the entire E1 measure (bottom panel).

have been applied to scale development and evaluation automatically, regardless of the type of construct being assessed, the structure of the data, or the intended purpose of the measure. Sadly, this approach has led to a proliferation—with no end in sight—of individual differences constructs and associated measures of varying and often low quality. IRT model application, however, requires researchers to, in advance, address a number of questions and meet several requirements. Much as in confirmatory factor analysis, the requirements of IRT models have forced researchers to be more circumspect when proposing constructs (an underlying latent variable) and in claiming that their item sets validly scale individuals on such constructs.

When considering an IRT application, the first question that needs to be asked is, "Is the construct and its measure (existing or proposed) consistent with such a model?" IRT models are latent variable modeling techniques (see Borsboom, 2005), and they make assumptions about item response data and the nature of the underlying construct (Bollen & Lennox, 1991). For example, latent variable models propose the existence of a continuous latent variable (serving as a proxy for a psychobiological trait), which, in some sense of the term, affects or causes variation in diverse behaviors (i.e., content-diverse items). These diverse behaviors (items) are correlated because of the common latent variable (i.e., the common cause). If behaviors or items are not content diverse, then correlations between them could be explained simply on the basis of semantic similarity, and there would be no need to propose a latent variable measurement model (see Tellegen, 1991).

Clearly, not all constructs that psychologists want to measure fit nicely under this latent variable rubric (see Bollen & Lennox, 1991). For example, Gough's "folk constructs" (Tellegen, 1993), assessed with the California Personality Inventory (Gough & Bradley, 1996), were developed to assess social perceptions, not latent traits. As such, applying an IRT model to these scales would be inappropriate. Moreover, many well-established scales were not developed using modern methods of scale analysis (e.g., confirmatory factor analysis) and are likely poor candidates for an IRT application.

For an example of a gold-standard measure that cannot be fit to an IRT model, see Reise, Horan, and Blanchard's (2011) analysis of the Social Anhedonia Scale (Eckblad, Chapman, Chapman, & Mishlove, 1982). Among other things, their analysis illustrates how having overly redundant item content increases interitem correlations and, thus, coefficient alpha, ultimately making the scale look better. IRT modeling also revealed multidimensionality, which makes measuring a common latent variable particularly challenging. Indeed, as we demonstrate, the BIS and NEO PI–R E1 data used in this chapter also illustrate the problem of content redundancy, which inflates traditional scale indices (e.g., coefficient alpha and interitem correlations) and ultimately greatly complicates meaningfully fitting an IRT model.

However, scales intending to measure conceptually narrow constructs or that are carefully constructed through the informed use of factor analysis (e.g., the Multidimensional Personality Questionnaire; Tellegen, 1995, 2003) are likely viable candidates for IRT modeling (e.g., Reise & Waller, 1990). As we underscore in the following, however, even carefully designed measures can present challenges because of the ever-present tension between wanting to scale individuals on one construct but needing to include content-diverse items to properly represent the construct. To be sure, obtaining an item set that affords measurement of a single latent variable that validly reflects what a diverse set of indicators have in common is both necessary and hard.

Assumptions of Unidimensional Item Response Theory Models

Unidimensional IRT models make three fundamental assumptions, two of which are interrelated. Next, we address these assumptions using these example data sets to illustrate various points. We then review the effects of failing to meet model assumptions on item parameter estimates and the identification of the latent variable. Ultimately, we argue that the degree to which the data meet the model's assumptions has profound consequences not only for the viability of IRT applications (e.g., computerized adaptive testing) but also for evaluating substantive hypotheses.

Monotonicity. The first assumption of unidimensional IRT modeling is that the relation between the latent variable and item response propensity is monotonically increasing; as trait levels increase, individuals score higher on the item. Some have referred to this as the *dominance response process assumption.* This assumption is necessary because logistic IRT models (Equations 6.3, 6.4, and 6.5) are parametric, monotonically increasing functions. (Note that this is not the case with nonparametric IRT models and unfolding models, which are more flexible in terms of shape of the IRC.)

The monotonicity assumption is easy to explore graphically using the rest-score function. For each item, a rest-score function is nothing more than a plot of the raw summed score (minus the item score) on the *x*-axis and the observed response proportions for each rest-score grouping on the *y*-axis. Typically, because there will seldom be enough people at each possible rest-score to compute a reliable proportion, rest-scores are often "binned" or grouped together (see the following examples). One can show that if item responses are monotonically increasing at the raw score level, then they must also be monotonically increasing at the latent variable level (Thissen & Orlando, 2001).

Rest-score curves and statistical tests of monotonicity are generated easily using the Mokken library (Van der Ark, 2007) available in the *R* 2.12 freeware statistical package (R Development Core Team, 2010). To illustrate, Figure 6.4 shows rest-score functions for three BIS items. Using the default bin-size selection option, the program grouped people into four rest-score groups: those with scores of 0–4, 5–7, 8–10, and 11–26. Notice that in the top graph (Item 6), the response proportions do not change much as a function of the raw scores, but in the middle graph (Item 17), response proportions increase nicely (high discrimination). In the bottom graph (Item 16), the largest monotonicity violation is shown, in which the response proportion for the highest group is .05 lower than the second highest rest-score group. Nevertheless, this is not a statistically significant violation (details about statistical testing are given in the manual). In fact, no BIS item significantly violated monotonicity.

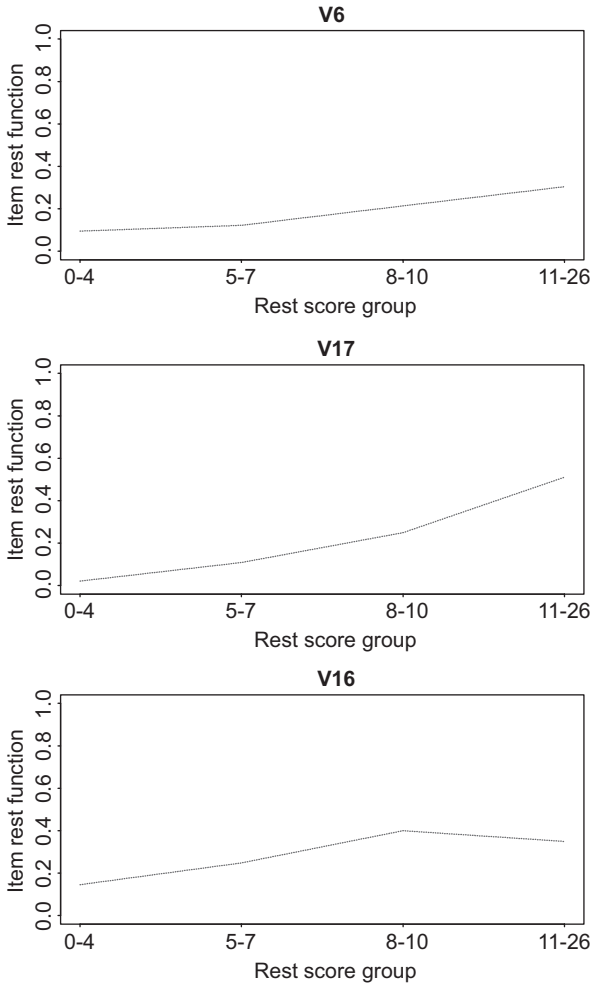

FIGURE 6.4. Rest-score curves for three BIS items (6, 17, and 16).

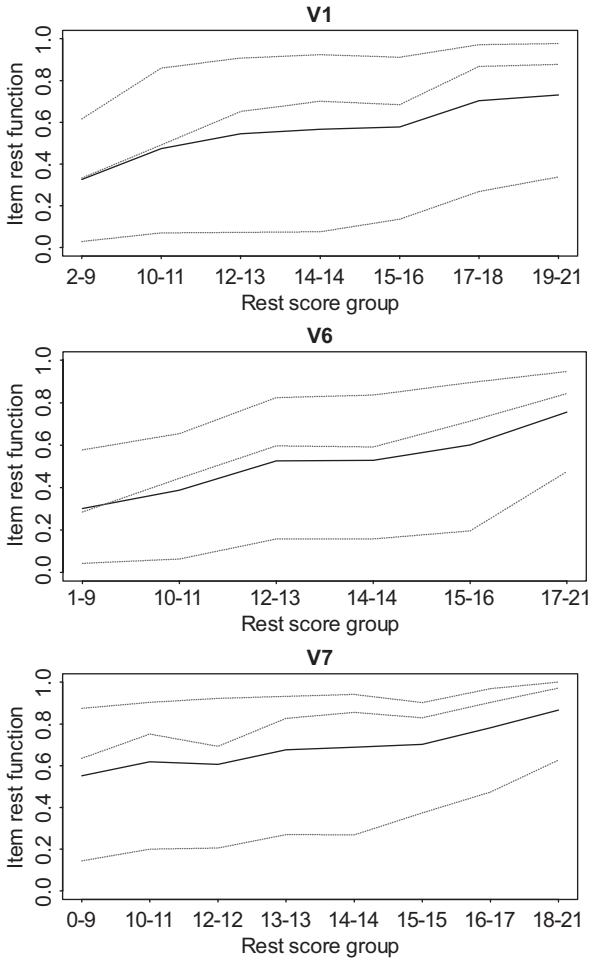

FIGURE 6.5. Rest-score curves for two E1 items (1 and 6).

Rest-score functions for polytomous items follow the same principle as those for dichotomous items, but they are a little trickier to interpret because there are $m = k - 1$ curves to inspect plus the curve for the mean response. Shown in the top portion of Figure 6.5, for example, is the rest-score function for NEO PI–R E1 Items 1 and 6, respectively. The solid line in the middle is the item mean score divided by the number of response categories for people falling in rest-score groups 1–9, 10–11, 12–13, 14, 15, 16–17, and 18–21. Clearly, the average item score is increasing as the rest-score increases, and there are no apparent violations for this item. The dotted lines represent response proportions. These curves should be increasing, and the space between them indicates conditional proportions. For example, the space from the top dotted line to the top of the figure is the conditional proportion responding zero, the space between the top two dotted lines is the conditional proportion responding 1, the space between the bottom two dotted lines is the conditional proportion responding 2, and the space from the bottom dashed line to the bottom of the figure is the conditional proportion responding 3. The steeper these curves are, the more discriminating the response categories. The relative spacing between the curves shows the relative response proportions; for Item 1, a response of 2 is most common, whereas for Item 6, the distribution of category responses is more evenly spread.

Although we need not show the rest-score curves for all E1 facet items, we note that regardless of their score on the measure, people tend to select either Category 2 or Category 3. Categories 0 and 1 are

seldom used, regardless of rest score. Much the same information could have been gathered from the category frequencies in Table 6.2. Nevertheless, evaluating monotonicity of category response both visually and statistically is still important to diagnose problematic category functioning. In these data, E1 Item 5 (bottom of Figure 6.5) was the worst-functioning item in the sense that three monotonicity violations, out of a possible 84 comparisons, were observed. The largest was a .04 decrease in proportions endorsed. Nevertheless, considered across the entire item, these minor violations are not statistically significant. As with the BIS, no items had statistically significant monotonicity violations.

Unidimensionality. Unidimensionality is the second assumption. Commonly applied IRT models assume that a single common latent variable accounts for all the common variance among the items. Stated differently, when controlling for a common factor, there should be no correlation among the items. On the surface, this requirement seems simple—most measures are designed to assess a single construct. However, except for the most conceptually narrow measures on which all items tap the same content theme, no real-world data set, strictly speaking, will be unidimensional. Because this issue is complex, before describing statistical approaches to the evaluation of unidimensionality, we first consider the concept of (uni) dimensionality and the reality of psychological measurement.

First, part of the problem in writing about and assessing dimensionality is that it is routinely misinterpreted in the literature. Dimensionality, whether unidimensionality or multidimensionality, is purely a statistical property of a correlation–covariance matrix; that is, only empirical data have dimensionality. Yet, it is common to find references to unidimensional or multidimensional constructs in the literature or to find statistical procedures (such as confirmatory factor analysis) that have demonstrated that a construct is unidimensional or multidimensional. Such verbiage regarding the dimensionality of a construct is very confusing, because a construct is a hypothetical entity proposed to explain some observed phenomenon; it is always one thing regardless of the content diversity or heterogeneity of its behavioral indicators.

A second complexity is that assessment writers often confuse the property of unidimensionality with the property that a single systematic factor influences scale scores. These ideas are related but distinct. *Dimensionality* refers to how many common latent factors influence item response behavior. We recognize that if a data set is truly unidimensional (one common factor), raw scores (or, by extension, IRT trait-level estimates) are unambiguously interpretable as reflecting a single latent variable and error. Even a unidimensional measure, however, can yield scores that reflect mostly error (e.g., if loadings are low). Moreover, even highly multidimensional measures can yield scores that predominantly reflect a single common latent variable. In fact, this phenomenon was noted long ago by Cronbach (1951) and continues to be emphasized by authors such as Gustafsson and Aberg-Bengtsson (2010) and Reise, Moore, and Haviland (2010). The important recognition here is that some multidimensional data sets can be fit quite nicely by a unidimensional IRT model; it depends entirely on the degree of multidimensionality and its structure.

Finally, it has long been recognized (e.g., Humphreys, 1970) that even if one could create a unidimensional measure or, more correctly, a measure that yielded unidimensional data, such an index may be undesirable because of a lack of content breadth. With this in mind, many researchers have noted that scales are typically constructed to contain item content representing the diverse manifestations of the construct (F. F. Chen, West, & Sousa, 2006) but, at the same time, predominantly measure one construct. This goal is accomplished by listing out, or empirically discovering, different aspects of a construct (i.e., different manifestations) and then writing sets of items (content parcels) that reflect this diversity (for a recent example, see Aluja, Kuhlman, & Zuckerman, 2010). Typically, scale developers will include three to five items to measure each of several distinct but correlated aspects of a construct.

This recognition—that many measures are designed explicitly to yield multidimensional data—is critically important because it influences one's

approach to evaluating dimensionality. Stated differently, the interest is not so much in statistical guidelines that tell one when a data set is or is not unidimensional or multidimensional or essentially unidimensional or has a strong first factor. Although those guidelines are the dominant indices used in the IRT literature, they raise and, to some extent, answer the wrong question. The question of interest, given multidimensionality caused by item content diversity, is "Can we still validly fit an IRT model?" That is, does multidimensionality bias or otherwise obfuscate attempts to measure a common dimension running among the items? Addressing this question requires a different statistical approach.

Assuming that item response data will typically not be purely unidimensional, IRT researchers have explored the robustness of IRT item parameter estimates to multidimensionality violations. We do not summarize this large literature here but note that a basic conclusion, and one reiterated in popular texts (Embretson & Reise, 2000), is that IRT item parameter estimates are reasonably robust if there is a strong common factor. In turn, several statistical approaches have arisen to explore whether data are unidimensional enough for application to IRT modeling. These approaches include inspection of eigenvalues from principal component or factor analysis (e.g., how large is the ratio of the first to second eigenvalue?), determining dimensionality through parallel analysis, comparing the fit of a one-factor confirmatory model against benchmarks (e.g., comparative fit index $> .90$), inspection of residuals after fitting a unidimensional IRT model, and many others.

We now consider some of these indices as applied to the example data sets. Starting with the BIS, the literature has shown that although the measure is ultimately used to yield a single raw score reflecting impulsivity, items were written to cover six different aspects of impulsivity. Given this content diversity, we do not expect the data to be perfectly unidimensional, but if we want to fit an IRT model, we need evidence of a common trait (and a strong one) running among the items.

To explore the number of common dimensions in the BIS data, we used the Psych library (Revelle, 2010) available in *R* freeware to conduct parallel analysis, inspect eigenvalues of a tetrachoric correlation matrix, perform cluster analyses from two to six clusters, and interpret minres factor solutions of between one and six (correlated) dimensions. Although the ratio of the first to second eigenvalues (7.3 to 3.4) is somewhat suggestive of a dominant first factor, from all analyses we concluded that there were between four and six discernable dimensions. Because six was the a priori hypothesis, for illustrative purposes we assume that six is the appropriate dimensionality. (If this were a research project we intended to publish, however, we would certainly explore and report on findings for several dimensional solutions.) Note that the finding of six factors is not necessarily inconsistent with unidimensionality; there could be one strong factor and several smaller "nuisance" dimensions caused by parcels of item content that are highly intercorrelated because they measure the same aspect of the construct.

The NEO PI–R E1 scale has only eight items and ostensibly measures only one narrow facet of extraversion—warmth—and hence there is no need for multi-item content parcels here. Thus, we expect the data from this scale to better conform to a "pure" unidimensional measure. Nevertheless, we conducted a parallel analysis, inspected eigenvalues of a polychoric correlation matrix, performed cluster analyses from two to three clusters, and interpreted minres factor solutions of between one and three (correlated) dimensions. Interestingly, parallel analysis based on principal component analysis suggested one component, but more important, a parallel analysis based on the common factor model suggested four factors. Yet, four factors for eight items is mathematically intractable, and when we explored a three-factor solution, only one item (7) loaded on the third factor. Although the ratio of the first to second eigenvalues was 3.5 to 1.0, suggesting unidimensionality, all other evidence pointed to one or two small secondary factors. We suspect that one or two small secondary "nuisance" factors may be caused by overly redundant item content (e.g., two items to represent interpersonal warmth vs. coldness, and two items for preference for verbal interaction with others). We explore this hypothesis more thoroughly next.

These analyses are useful in exploring the dimensionality of an item response matrix, determining the relative strength of a single factor, and checking whether a priori hypothesized content dimensions manifest in real data. Yet, all these approaches are only indirect methods that do not necessarily tell a researcher what he or she needs to know, namely, the extent to which item parameter estimates are biased by multidimensionality in the data. In other words, the real question is do the item parameters reflect what is in common among the items or are they biased by multidimensionality? To better address this question, and others, several researchers have suggested that a bifactor model be used as a framework for exploring the unidimensionality assumption in IRT (Reise, Cook, & Moore, in press; Reise, Morizot, & Hays, 2007).

A bifactor model is a factor structure in which all items load on a general dimension (the target trait) and one so-called "group" factor. The general dimension and the group factors are all orthogonal. This model, it has been argued, is more consistent with the view that item response data measure a single common dimension but also contain secondary common factors caused by content parcels as described earlier. These group factors that emerge because of content parcels have historically been referred to as nuisance dimensions (they interfere with the measurement of the construct of interest).

As argued in Reise, Cook, and Moore (in press) and Reise, Moore, and Maydeu-Olivares (2011), an exploratory bifactor model is useful to address several important questions. First, by inspecting the size of the loadings on the general factor, one can judge whether the items all reflect a single common dimension and, if so, how strongly. Note that in the bifactor model, the general factor is, in a sense, free of the contamination caused by multidimensionality. Second, inspection of the factor loadings on the group factors shows the degree to which items are being influenced by secondary common factors, potentially preventing any attempt to fit a unidimensional model. Third, one can compare the loadings on the general factor of the bifactor model with the loadings from a unidimensional model; if not dramatically different, it would be hard to argue that multidimensionality made much difference. Fourth, and more important, the bifactor allows the computation of two highly useful statistics: (a) the proportion of common variance due to the general factor—a direct index of unidimensionality—and (b) coefficient omega hierarchical (ω_h; McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005), which reflects the percentage of variance in scores due to the general factor.[1]

To illustrate, we again used the Psych library available in the *R* statistical package to estimate an exploratory bifactor model, specifically the Schmid–Leiman orthogonalization (Schmid & Leiman, 1957), and to compute coefficient omega hierarchical. For the BIS, we specified one general and six group factors and for the NEO PI–R E1 data, we specified one general and two group factors. Note that our analyses were based on polychoric (E1) or tetrachoric (BIS) correlations, not Pearson correlations. Thus, the interpretation of the coefficient alpha and coefficient omega hierarchical statistics to be reported on must be considered in this light.

The results for the BIS are shown in Table 6.3 (loadings of .30 and higher). Starting with the bifactor results, inspection of the loadings on the general factor makes it clear that the scale contains many items that do not measure the general trait (impulsivity). Although much the same could have been determined from a one-factor model (third column), Items 10 and 30 would mistakenly have been included in the measure. Moreover, comparison of the loadings on the group factor with the loadings

[1]To interpret coefficient omega hierarchical, one must be mindful of what is being analyzed. Coefficient omega is the sum of loadings on the general factor, squared (not the sum of the squared loadings), divided by the sum of the variance–covariance matrix of item responses. By the variance sum law, the sum of the variance–covariance matrix is equal to the variance of raw scores. There are three situations: (a) If the covariance matrix is factor analyzed, then coefficient omega hierarchical is the percentage of raw score variance explained by the general factor and is similar to a correlation between the observed scores and a single underlying latent variable; (b) if the Pearson correlation matrix is factor analyzed, coefficient omega hierarchical is the percentage of summed standardized item score variance explained by a general factor, which is analogous to the distinction between alpha (computed on a covariance matrix) and standardized alpha (computed on a correlation matrix); and (c) if one begins with a tetrachoric or polychoric correlation matrix, the interpretation of omega hierarchical is more complicated because the denominator does not refer to the variance of either summed raw scores or summed standardized item scores. Instead, the denominator term is the variance of the underlying item response propensities. This does not defeat its purpose—high values of omega hierarchical mean that a single latent trait is the dominant influence on item responses and ultimately latent trait-level estimates.

TABLE 6.3

Factor Loadings in a One-Factor Model and a Bifactor Model for the Barratt Impulsivity Scale Version 11 (Impulsivity) Data

				Group factors					
Item	**Abbreviated content**	**1-Fac**	**General factor**	**1**	**2**	**3**	**4**	**5**	**6**
12	Thinks carefully	.60	.41	.80					
20	Thinks steadily	.60	.40	.65					
9	Concentrates	.67	.52	.52	.38				.30
15	Thinks about complex problems			.51		.34			
1	Plans carefully	.51	.33	.49					
8	Self-controlled	.61	.44	.32					
28	Restless theater–lectures	.52	.45		.67				
11	Squirmy plays–lectures	.46	.39		.61				
18	Gets bored with complex problems	.48	.41		.44	.40			
14	Speaks without thinking	.60	.48		.30				
26	Extraneous thoughts pop into head	.55	.45			.54			
6	Racing thoughts	.47	.38			.54			
21	Changes where lives	.30				.45			
24	Changes activities	.44	.34			.35			
27	Present oriented					.31		.36	
19	Acts spur of the moment	.67	.63				.76		
17	Acts on impulse	.80	.67				.57		
13	Plans for future job security	.42	.31					.71	
10	Has a regular savings plan	.33						.53	
30	Future oriented	.38						.48	
25	Spends over budget	.46	.33					.39	
16	Changes jobs							.37	
7	Long-term planning of trips	.40						.36	
5	Doesn't pay attention	.44	.44						.88
23	Thinks of one thing at a time								.37
2	Acts without thinking	.68	.53						
22	Impulsive buying	.48	.33						
3	Quick decision making								
4	Happy-go-lucky								
29	Likes to work on puzzles								

Note. 1-Fac is one factor minres factor analysis. Loadings below .30 are not shown.

on the general factor in the bifactor model shows that the loadings in the unidimensional model are highly inflated because of multidimensionality. Once that multidimensionality is controlled for, as in the bifactor model, the items appear to be relatively much weaker indicators of impulsivity.

The bifactor group factor loadings show that the data are contaminated not so much by multiple interpretable content dimensions, but by doublets and triplets of items, which in turn manifest as group factors. Consider Items 12 and 20, which are essentially the same item, one with the word *careful* and the other with the word *steady*. Group Factor 2 is defined by the item pair asking about whether one is restless or squirmy at social functions that require long periods of attention. Coefficient alpha is .87 (computed using tetrachorics), which meets or exceeds acceptable internal consistency standards. Yet, coefficient omega hierarchical is only .51 (computed using tetrachorics). The difference in these two statistics provides an estimate of the degree to which coefficient alpha overestimates the precision

of the measure resulting from multidimensionality. Coefficient omega hierarchical is the proper statistic to use if a researcher is interested in the degree to which scores reflect variation on a single common factor (i.e., the general factor in a bifactor model). Finally, given that the percentage of common variance due to the general factor is only 32%, we conclude that the data provide little evidence of a strong common factor. This instrument appears to be the type of measure that is better represented by a set of subscales.

The analogous results for the NEO PI–R are shown in Table 6.4. In the second column are the loadings from a one-factor model, and in the remaining columns are the loadings from the general bifactor model and two group factors. As with the BIS, a comparison of the loadings on the one-factor solution with the loadings on the general factor shows that loadings are inflated around .10 in the one-factor solution. Unlike the BIS, however, all E1 items show at least modest loadings on the general factor in the bifactor, and a few items have strong loadings. Inspection of the group factor loadings shows the presence of two item doublets: Group Factor 1 contains Items 2 and 5, and Group Factor 2 contains Items 3 and 4. Interestingly, although there are only two group factors, the percentage of common variance due to the general factor is only 56%. This suggests that the group factors have almost as much influence on the reliable portion of the resulting summed scores as the target dimension. Finally, coefficient alpha is .81 (computed using polychoric correlations) and coefficient omega hierarchical is .56 (with polychoric correlations). As with the BIS, it is clear that alpha is highly inflated by multidimensionality.

TABLE 6.4

Factor Loadings in a One-Factor Model and a Bifactor Model for the Revised NEO Personality Inventory E1 Scale Data

		General	Group factors	
Item	1-Fac	factor	G	G
1	.53	.43		
2	.66	.51	.47	
3	.63	.61		.64
4	.62	.53		.45
5	.76	.65	.67	
6	.57	.47		
7	.47	.35		
8	.47	.36		

Note. 1-Fac is one factor minres factor analysis. Loadings below .30 not shown.

Local independence. Third, and very much related to the second, is the assumption of local independence. Local independence and unidimensionality are intertwined but distinct concepts. Technically speaking, unidimensionality exists when item responses are locally independent (e.g., uncorrelated) after controlling for a single common factor. Thus, claiming that a data set is unidimensional is the same as claiming that responses are locally independent after extracting a single factor. However, there are situations in which local dependence (LD) can occur, but we would not want to claim multidimensionality. For example, in verbal tests, sets of items (testlets) are often attached to a given reading comprehension passage. These items are often more highly intercorrelated with each other than they are with the remaining items. This situation causes a LD violation, but we would certainly not claim a second dimension, although perhaps, technically speaking, we could if one were identified.

A second type of LD violation occurs when a noncognitive measure contains a near duplicate pair of items ("I'm happy almost all the time" and "I'm happy much of the time"). Again, such items will be correlated with each other beyond what can be explained by a single common factor (because the items share the common factor, depression, and a group content factor, happiness). This is a LD violation but not necessarily a violation of unidimensionality. Pairs of items with LD violations are important to identify because they can inflate psychometric indices such as coefficient alpha as well as prevent researchers from correctly modeling the underlying common latent variable, as described in more detail next.

Several statistics are designed to identify LD violations (Yen, 1993), but the W. Chen and Thissen (1997) approach included in mvIRT (Cai et al., 2011) is what we use here. We do not summarize and comment on all of the technical details of this

statistic, but suffice it to say that it is a standardized residual and that one is to look for large positive numbers. *Large* remains ill defined in the literature, but values greater than 5 and most certainly those greater than 10 should be taken as severe LD violations.

In these data, seven item pairs were identified on the BIS that had LD statistics greater than 10: 1 and 13 ("plans"), 15 and 12 ("thinks"), 21 and 16 ("changes"), 26 and 6 ("extraneous" or "racing thoughts"), 28 and 11 ("restless" or "squirms"), 29 and 15 ("puzzled" and "complex problems"), and 30 and 27 ("future"). The results for the NEO PI–R E1 scale are unsurprising given the previous bifactor results; Items 2 and 5 ($z_{LD} = 15.20$) and Items 3 and 4 ($Z_{LD} = 16.70$) result in severe violations of the local independence assumption. Smaller violations were identified between Items 5 and 3 ($Z_{LD} = 6.60$) and between Items 8 and 7 ($Z_{LD} = 7.90$). Clearly, there is redundancy between the LD analysis and the bifactor analyses shown previously. The LD analysis, however, identifies item pairs, whereas the bifactor modeling more readily identifies content-thematic groups of items that violate local independence. Both types of analyses should be conducted.

Why Assumptions Are Important

At this juncture, we trust it is clear that the advantages and applications of latent variable models are only viable to the degree that the data conform to model assumptions. IRT model applications, such as differential item functioning analysis, scale parameter linking, and computerized adaptive testing, rest entirely on the properties of item and person invariance, which, in turn, can be interpreted only when the data meet IRT model assumptions. That said, if an IRT application truly required strictly meeting all assumptions, there would be no unidimensional IRT applications. How does one proceed in the face of this? To answer this question, we need to be clear on why assumptions are needed and the consequences of violation.

Monotonicity is the easiest to understand and evaluate. It is also the most likely to be satisfied easily, given that scales include items with generally acceptable item–test correlations. The reason monotonicity is required is that logistic models will force a positively increasing relationship between the latent variable and item response proportions. If the data do not conform to this assumption, then the parameters of logistic IRT models have no valid interpretation, and alternative models (e.g., nonparametric or unfolding) should be considered.

Unidimensionality and the effects of violating it are complicated. As noted, except for the most conceptually narrow of measures (very homogeneous item content), we do not expect unidimensionality to be satisfied. For this reason, IRT proponents have emphasized the notion of a strong common trait or strong first factor, and it can be shown that in many real-life situations, item parameter estimates will be drawn toward that strong first factor (Drasgow & Parsons, 1983). Nevertheless, in considering unidimensionality, there really are two issues. First, do the estimated item parameters reflect the common dimension running among the items or are they distorted by multidimensionality? Second, and related to the first, is the chosen common latent variable identified correctly, or is it somehow distorted by multidimensionality (i.e., pulled toward a group content factor)?

A concern we have expressed throughout this chapter is that the standard approaches to evaluating "unidimensional enough" do not directly address these questions, nor are they helpful in understanding what revisions a test author might make to create a measure more amenable to IRT analysis. For example, there is no citable and defensible benchmark size of the ratio of the first to second eigenvalues that ensures the appropriateness of IRT model application. The same holds for more sophisticated approaches, such as inspecting practical fit indices derived from confirmatory factor modeling. What exactly, for example, would a "robust" comparative fit index of .86 and a root-mean-square error of approximation of .064 mean in terms of the accuracy of the estimated item parameters? Very little.

In place of these unidimensional-enough indices, we suggest comparison modeling (Reise, Cook, & Moore, in press). In comparison modeling, one assumes that the measure does assess a common latent variable but that there is multidimensionality caused by clusters of items with similar content. One then estimates the number of nuisance dimensions

or group factors caused by this content diversity and fits an exploratory bifactor model.[2] Judgments about the reasonableness of IRT modeling are based on the inspection of loadings (especially loadings on the general factor compared with the loadings in a unidimensional model). One then computes the percentage of common variance due to the general factor (not total variance) and coefficient omega hierarchical (to judge whether scores derived from the measure can be viewed as reflecting primarily a single source).

To the degree that these statistics are high, items load strongly on the general factor and have loadings on the general factor similar to loadings in a one-factor model, one can more comfortably argue for the appropriateness of undimensional IRT model application. The NEO PI–R E1 scale is a good example of a measure that appears to be a viable candidate for IRT modeling if two of the items (one from each doublet) involved in the LD violations were deleted. The BIS, however, is not. Many items do not load on the general factor, and those that do have modest loadings or load higher on group factors. In short, there does not appear to be a common strong impulsivity latent variable to model. Moreover, the instrument contains many doublets or triplets of similarly worded items. When this happens, we cannot be confident that the IRT parameters are valid or determine whether they are biased positively by the inflated relation between two items that are essentially the same item asked twice.

Of course, some may argue that we have simply replaced one set of ambiguous indices (eigenvalue ratios) with an equally problematic alternative. Although we readily concede that this alternative has many challenges (e.g., determining the number of group factors), it allows one to make a more informed judgment about the effects of multidimensionality on fitting a unidimensional model. Moreover, it yields two readily interpretable indices. For example, the percentage of common variance is a direct and easily interpretable index of unidimensionality (how much common variance is explained by a single factor), and coefficient omega hierarchical assists one in judging the degree to which the latent variable truly reflects a single common variable that runs among the items. Finally, although not a topic in this chapter, our approach places the researcher in a good position to consider alternative multidimensional IRT models (Reckase, 2009). Note that IRT methodology other than the bifactor model can be applied to higher dimensioned data (e.g., IRT models for two or more possibly correlated latent variables). We present a bifactor model application only, for only the bifactor multidimensional IRT model allows researchers to maintain the goal of using an instrument to measure one common latent variable.

Local independence violations are important for the same reasons violations of unidimensionality are important (Steinberg & Thissen, 1996). If allowed to stay in the measure, local dependencies can result in distorted parameter estimates and misidentification of the latent variable. Another point emphasized in the LD literature is that violations cause standard errors to be biased low. That is, by including a LD violation, researchers credit themselves for asking the same question twice. Most important, an LD violation can lead to an IRT slope parameter (or factor loading) estimate that is too high (see unidimensional solutions in Tables 6.3 and 6.4). The reason this occurs is that the latent variable is, in a sense, "pulled" toward the items with the highest item intercorrelations. For example, although not shown, when we fit an IRT model to a fears scale, the two content-redundant fear-of-the-dark items (which were correlated around .85) had slope parameter estimates in the 2PL model greater than 6 (ridiculously high). When one of the LD items was removed, the slope parameter estimate for the remaining item went back down to a more reasonable value, and consequently, the latent variable better reflected fears rather than the more specific construct fear of the dark.

Finally, now that we have commented on the consequences of model violations, we argue more generally that having a valid measurement model is of paramount importance for substantive researchers. This argument follows that of Little, Lindenberger, and Nesselroade's (1999) explorations of parceling in latent variable models. They explored

[2]Throughout, we used a Schmid–Leiman procedure, but there are attractive alternatives such as targeted bifactor rotations (Reise, Moore, & Maydeu-Olivares, 2011) and analytic bifactor rotations (Jennrich & Bentler, 2011).

the consequences of different ways of specifying a measurement model for determining a latent variable's correlation with criterion measures. In other words, they asked whether measuring the target construct correctly really matter in terms of estimating validity coefficients. Their results clearly showed that (a) if raw sum scores are used to represent a construct, validity coefficients can be wildly biased and misleading and, more important, (b) if latent variable modeling is used and the target construct is properly measured, validity coefficients can be recovered with high accuracy. An analogous set of arguments apply to IRT modeling. In contexts in which it is critical to represent a latent variable correctly, having an IRT model in which the data meet the assumptions is clearly necessary.

CONCLUSION

A primary goal of this chapter was to clarify the assumptions underlying IRT modeling and offer ways of more thoughtfully scrutinizing one's data because, clearly, IRT preparatory work and modeling can improve existing measures and guide researchers in the development of new measures. IRT applications, however, have been somewhat haphazard, much as applications of other newer "sophisticated" and "promising" statistical methods (e.g., structural equation models, multilevel modeling, and latent growth curve analysis). To avoid the "have hammer, must nail things" phenomenon, we recommend that when considering IRT applications, researchers (and journal editors) base decision making (at the scale and study levels) not on the conventional acceptability ("rules-of-thumb") standards but rather on evaluations of modern techniques for examining an instrument's structure and fitting alternative exploratory multidimensional models, such as a bifactor. Research articles on the psychometrics of instruments should not read like advertisements; they should be thoughtful and complete presentations of a measure's strengths and weaknesses with respect to a measurement model.

References

Aluja, A., Kuhlman, M., & Zuckerman, M. (2010). Development of the Zuckerman-Kuhlman-Aluja Personality Questionnaire (ZKA-PQ): A factor/facet version of the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ). *Journal of Personality Assessment, 92*, 416–431.

Barratt, E. S. (1959). Anxiety and impulsiveness related to psychomotor efficiency. *Perceptual and Motor Skills, 9*, 191–198. doi:10.2466/pms.1959.9.3.191

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305–314. doi:10.1037/0033-2909.110.2.305

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511490026

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . PROMIS Cooperative Group. (2007). The patient-reported outcomes measurement information system (PROMIS). *Medical Care*, *45*(5, Suppl 1), S3–S11. doi:10.1097/01.mlr.0000258615.42478.55

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality-of-life. *Multivariate Behavioral Research, 41*, 189–225. doi:10.1207/s15327906mbr4102_5

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289. doi:10.3102/10769986022003265

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI–R) and NEO Five-Factor Inventory (NEO FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189–199. doi:10.1177/014662168300700207

Eckblad, M. L., Chapman, L. J., Chapman, J. P., & Mishlove, M. (1982). *The Revised Social Anhedonia Scale* [Unpublished test]. Madison: University of Wisconsin.

Embretson, S. E., & Reise, S. P. (2000). *Psychometric methods: Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Gough, H. G., & Bradley, P. (1996). *CPI manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97–121). Washington, DC: American Psychological Association. doi:10.1037/12074-005

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 23–32). Seattle: University of Washington.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*, 537–549. doi:10.1007/s11336-011-9218-4

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*, 192–211. doi:10.1037/1082-989X.4.2.192

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

Morizot, J. M., Ainsworth, A. T., & Reise, S. P. (2007). Towards modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). New York, NY: Guilford Press.

Multivariate Software, Inc. (2010). *mvIRT—A user-friendly IRT program*. Encino, CA: Author.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. Thousand Oaks, CA: Sage.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51*, 768–774. doi:10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

R Development Core Team. (2010). *R: A language and environment for statistical computing, reference index version 2.12.1*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. doi:10.1007/978-0-387-89976-3

Reise, S. P. (2010). The emergence of item response theory (IRT) models and the Patient Reported Outcomes Measurement Information System (PROMIS). *Austrian Journal of Statistics, 81*, 93–103.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*, 95–101. doi:10.1111/j.0963-7214.2005.00342.x

Reise, S. P., Cook, K. F., & Moore, T. M. (in press). A direct modeling approach for evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. R. Reise & D. Revicki (Eds.), *Handbook of item response theory and patient reported outcomes*. London, England: Taylor & Francis.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI–R. *Assessment, 7*, 347–364. doi:10.1177/1073191 10000700404

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93–103. doi:10.1207/S15327752JPA8102_01

Reise, S. P., Horan, W. P., & Blanchard, J. J. (2011). The challenges of fitting an item response theory model to the Social Anhedonia Scale. *Journal of Personality Assessment, 93*, 213–224. doi:10.1080/00223891.2011.558868

Reise, S. P., & Moore, T. M. (2012). An introduction to item response theory models and their application in the assessment of noncognitive traits. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 699–721). Washington DC: American Psychological Association. doi:10.1037/13619-037

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544–559. doi:10.1080/00223891.2010.496477

Reise, S. P., Moore, T. M., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement, 71*, 684–711. doi:10.1177/0013164410378690

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19–31. doi:10.1007/s11136-007-9183-7

Reise, S. P., & Waller, N. G. (1990). Fitting the two parameter model to personality data: The parameterization of the Multidimensional Personality Questionnaire. *Applied Psychological Measurement, 14*, 45–58. doi:10.1177/014662169001400105

Reise, S. P., & Waller, N. (2009). Item response theory and clinical measurement. *Annual Review of*

Clinical Psychology, 5, 27–48. doi:10.1146/annurev.clinpsy.032408.153553

Revelle, W. (2010). *Package "psych": Procedures for psychological, psychometric, and personality research (R Package Version 1.0–68)*. Retrieved from http://personality-project.org/r and http://personality-project.org/r/psych.manual.pdf

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores *Psychometrika Monograph Supplement*, *34,* 100–114.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53–61. doi:10.1007/BF02289209

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81–97. doi:10.1037/1082-989X.1.1.81

Tellegen, A. (1991). Personality traits: Issues of definition, evidence and assessment. In D. Cicchetti & W. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (pp. 10–35). Minneapolis: University of Minneapolis Press.

Tellegen, A. (1993). Folk concepts and psychological concepts of personality and personality disorder. *Psychological Inquiry, 4*, 122–130. doi:10.1207/s15327965pli0402_12

Tellegen, A. (1995). *Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.

Tellegen, A. (2003). *MPQ scales*. Unpublished manuscript, University of Minnesota.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Erlbaum.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*. Retrieved from http://www.jstatsoft.org

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133. doi:10.1007/s11336-003-0974-7

CHAPTER 7

ITEM ANALYSIS

Randall D. Penfield

Psychological assessments come in a multitude of forms and lengths and are used to measure a broad range of constructs. Despite their varied attributes, all psychological assessments share the property of being composed of a series of items, tasks, or questions to which an individual provides a response. Simply stated, items are the building blocks of psychological assessments. An individual's responses to the items on an assessment are used to make inferences about the individual's level of the psychological trait being measured, most commonly through the creation of a score reflecting the individual's level of the trait.

Given that items serve as the foundational components of psychological assessments, the quality of the assessment scores (i.e., reliability and validity) is dictated by the properties of the items making up the assessment. Good items lead to good-quality scores and bad items lead to bad-quality scores. But how does one determine whether a particular item is good or bad? This is the purpose of item analysis, a process by which the properties of items are evaluated with the goal of determining (a) which items are and which items are not making an acceptable contribution to the quality of the scores generated by the assessment and (b) which items should be revised or removed from the assessment altogether. This chapter provides a conceptual overview of item analysis and describes statistical methods used to conduct item analyses.

OVERVIEW OF ITEM ANALYSIS AND ITS USE IN ASSESSMENT DEVELOPMENT

Let me begin this discussion of item analysis with the simple assertion that the purpose of any assessment is to make inferences about the respondent's level of the psychological trait of interest. For simplicity, I refer to the psychological trait measured by a particular assessment as the *target trait*. Because an assessment is a collection of items, it follows that each item of the assessment is intended to contribute to this purpose by providing information about the respondent's level of the target trait. The amount of information provided by an item is determined by the extent to which the response to the item contributes to an understanding of the respondent's level of target trait. Any item providing a negligible amount of information is undesirable because such an item expends valuable respondent time while contributing little to the quality of the inference of a respondent's target trait. As a result, the goal for any assessment is to contain only items that provide a high level of information concerning the target trait for the population of interest. An item analysis is used to quantify the information provided by each item and to identify faulty properties of items contributing negligible information in the hopes of revising such items appropriately.

Two Components of Item Information: Discrimination and Difficulty

The amount of information provided by an item for the intended population is determined primarily by two psychometric properties: (a) item discrimination and (b) item difficulty. The property of item discrimination concerns how well the item's response categories distinguish between individuals having different levels of the target trait. The more the response categories differentiate (or discriminate)

DOI: 10.1037/14047-007
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

between individuals with different trait levels, the more information the item provides about the target trait level. Although item discrimination addresses how much information the item provides about the target trait, it does not address which levels of target trait are informed by the item's responses. Items differ with respect to the range of the target trait informed by the item; some items provide information about respondents having low levels of the target trait, and some items provide information about respondents having high levels of the target trait. This is where item difficulty comes into play: Item difficulty concerns the portion of the target trait continuum for which the item provides information. Items that have high difficulty provide information about respondents with high target trait levels, and items that have low difficulty provide information about respondents with low target trait levels.

To illustrate the concepts of item discrimination and difficulty, I present a diagram in Figure 7.1 for which (a) the horizontal axis represents the target trait continuum, (b) the vertical axis represents information concerning a respondent's target trait level, and (c) arrows represent items. Let us consider an assessment containing four items (the particular format of the items is unimportant for this illustration), each of which has a corresponding arrow in the figure. For each item, the height of the arrow reflects the item's discrimination, and the location of the arrow reflects the item's difficulty. Thus, higher arrows correspond to more information, and the location of the arrow reflects the target trait level for which the item provides information.

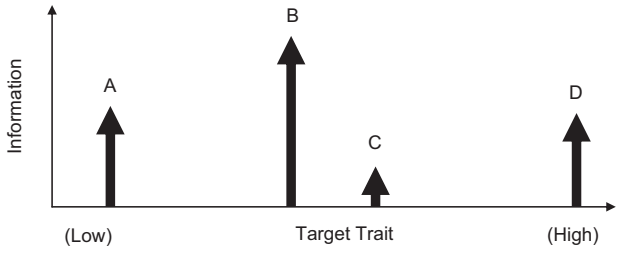

FIGURE 7.1. Diagram depicting item discrimination (height of arrow) and difficulty (location of arrow) for four items labeled A (low difficulty, moderate discrimination), B (moderate difficulty, high discrimination), C (moderate difficulty, low discrimination), and D (high difficulty, moderate discrimination).

For example, Item A provides information primarily targeting individuals who have a low target trait value (low difficulty), and the amount of information can be viewed as moderate (moderate discrimination). Similarly, Item B provides high information targeting individuals who have a moderate target trait value (high discrimination, moderate difficulty), Item C provides relatively low information targeting individuals who have a moderate target trait value (low discrimination, moderate difficulty), and Item D provides moderate information targeting individuals who have a high target trait value (moderate discrimination, high difficulty).

Using discrimination and difficulty can aid the assessment developer in determining which items require revision or removal. In the example shown in Figure 7.1, the assessment developer would be well served by taking a closer look at Item C for potential revision or removal. In addition, the assessment developer should take note of the relatively extreme difficulty (locations) of Items A and D, because these items provide information about respondents with very low (Item A) and very high (Item D) levels of the target trait; it is then necessary to consider whether these extreme levels of target trait are represented in the population of interest.

Using Item Analysis in the Assessment Development Process

The properties of item discrimination and difficulty play fundamental roles in determining the reliability and validity of the final scores generated by the assessment. Without adequate discrimination across the items of the assessment, the assessment cannot generate scores that are valid and reliable, regardless of the target trait levels of the individuals in the population. Without an appropriate level of difficulty across the items of the assessment, the assessment cannot generate scores that are valid and reliable at particular points of the target trait continuum. It should come as no surprise, then, that item analyses play a fundamental role in the assessment development process. Item analysis guides the assessment developer in identifying the best possible set of items within the practical constraints of the assessment development process.

The results of an item analysis are commonly used at several points during the assessment

development process. Item analysis is typically first encountered in the context of formal pilot testing, whereby the items of the instrument are administered to a sample of individuals to generate information concerning the psychometric properties of the items and the assessment as a whole. At this stage of assessment development, the results of an item analysis provide rich information concerning which items need to be revised or removed and whether additional items are needed to meet the goals of the assessment. In evaluating the results of an item analysis, the assessment developer must balance what is ideal against what is possible or reasonable within the limitations of the development process. In many instances, high discrimination for all items is an unrealistic outcome, as is attaining item difficulties that align exactly with a predetermined range of values. Thus, determining which items to retain or remove from a pool of piloted items must be interpreted within the context of the practical constraints facing the item developer.

Modifications to the item pool on the basis of the results of an item analysis must also be made with consideration of the content domain underlying the assessment. Assessments are constructed with the intent of generating information about a respondent's level of the target trait of interest, but target traits are typically operationalized according to a domain of content, attributes, or behavior. For example, the trait of risk taking may have a content domain that is operationalized according to numerous components, just a few of which include risk taking in relation to one's health, financial status, and social standing (Blais & Weber, 2006), and each of these components can be subdivided further with respect to specific behaviors (e.g., health risk taking can include going white water rafting or driving a car without wearing a seatbelt). The items of an assessment provide a sample of the individual behaviors or attributes contained within the intended content domain. In removing items from the assessment, one must always take into consideration the content domain or domains intended to be sampled by the instrument. At times, removing particular items can reduce the representation of a particular content domain to an unacceptable level, and thus either the item must be retained (despite less than ideal discrimination) or other items must be created to compensate for the lack of representation in the desired domain. This process is not one size fits all, and the extent to which content representation is affected by item removal through the item analysis process will depend on the particular assessment.

In addition to conducting an item analysis during the pilot testing phase of instrument development, item analysis is commonly used after any large administration of an assessment to make certain that the items are working appropriately for the intended population, and any poorly performing items can potentially be removed before estimating the respondents' level of the target trait. However, as stated earlier, the removal of items must be conducted with consideration of the impact on the assessment content.

Last, the results of an item analysis provide important information concerning the validity of the scores generated by the assessment. The process of validation incorporates the collection of several forms of evidence of validity, including content-based evidence, criterion-based evidence, response-process evidence, and internal structure evidence (American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education, 1999). Item analysis provides useful information related to internal structure evidence of validity, which concerns the extent to which the components (e.g., items) of the assessment are related to one another in a manner that is consistent with the intended target trait structure (Loevinger, 1957). Item discrimination addresses the extent to which each item generates information concerning the target trait (i.e., the extent to which all items measure a common target trait). Item difficulty addresses the level of target trait about which each item provides information; the relative difficulty of the items should be consistent with that expected by the content of the items.

ITEM DISCRIMINATION AND DIFFICULTY: A CLOSER LOOK

The previous section introduced the concepts of item discrimination and difficulty as the primary determinants of the information an item provides

about respondents' target trait levels. I now expand on this introduction, providing a more comprehensive description of what item discrimination and difficulty really are. This expanded description will prove valuable in understanding the content of later sections of this chapter describing statistical methods used to quantify item discrimination and difficulty.

Assumption of Monotonicity

In all discussions of item discrimination and difficulty that follow, I make the assumption that the scoring of the item is conducted in a monotonically increasing fashion, such that higher score levels on an item are associated with successively higher levels of the target trait. For example, for an item with score levels of 1, 2, and 3, one assumes that respondents who score a 3 tend to have a higher target trait level than those who score a 2, who in turn tend to have a higher target trait level than those who score a 1. It is possible for items to be initially coded in a nonmonotonically increasing fashion (e.g., lower score categories correspond to higher levels of the target trait), as might be the case if one is using the level of social activity (e.g., 1 = *never,* 2 = *sometimes,* 3 = *often*) as a measure of social anxiety (lower levels of social activity are expected to correspond to higher levels of social anxiety). However, these initial response categories would need to be recoded appropriately before analysis so that higher item score levels are associated with higher levels of the target trait. In the example of social anxiety, this would require a recoding of the responses as 1 = *often,* 2 = *sometimes,* and 3 = *never*. Failure of items to be monotonically increasing will cause item discrimination and difficulty to assume nonsensical values (as described next) and will threaten the validity and reliability of the target trait estimates generated by the assessment.

Item Discrimination

The initial introduction to item analysis described item discrimination in terms of the amount of information an item provides about the respondent's level of a target trait, which is determined by the extent to which the response categories of the item discriminate, or differentiate, between individuals in different portions of the target trait continuum. An understanding of discrimination can be deepened using a relatively modern psychometric concept of the item response function (IRF). To describe the concept of the IRF, consider a hypothetical item having three ordered response categories coded as 1, 2, and 3. This item could be a rating scale item whereby the values of 1, 2, and 3 reflect levels of agreement (*disagree, neutral, agree*), magnitude (*none, some, a lot*), correctness (*completely incorrect, partially correct, completely correct*), or any one of many other ordinal quantifications. For this item, one can consider the probability (or chance) of observing each possible response category (1, 2, or 3) as a function of the target trait level. An example of this is presented in Figure 7.2 for two different items (A and B). In Figure 7.2, the probability of observing each response category is displayed at each level of target trait, where target trait is on a standardized metric for which a value of zero can be viewed as moderate, high negative values represent relatively low levels of the target trait, and high positive values represent relatively high levels of the target trait. Notice that there is a separate line (or function) for each response category (i.e., a line for 1, a line for 2, and a line for 3), and each of these lines is an IRF. Each IRF can range in height between 0 (no chance of observing the response category) and 1 (100% chance of observing the response category).

The nature of the IRFs shown in Figure 7.2 defines the item's discrimination; high discrimination occurs when the response categories have a high chance of occurring for distinct ranges of target trait. For example, the item in the top portion of Figure 7.2A has high discrimination because the item response provides nearly unambiguous information concerning the portion of the target trait continuum to which the respondent belongs: Individuals scoring a 1 are almost certain to have a target trait level that is less than −1.2 (i.e., individuals who have a target trait level less than −1.2 are almost certain to score a 1), individuals having a response of 2 are almost certain to have a target trait level that is between −0.8 and 0.8, and individuals having a response of 3 are almost certain to have a target trait level that exceeds 1.2. Thus, the outcomes associated with this item are very effective at

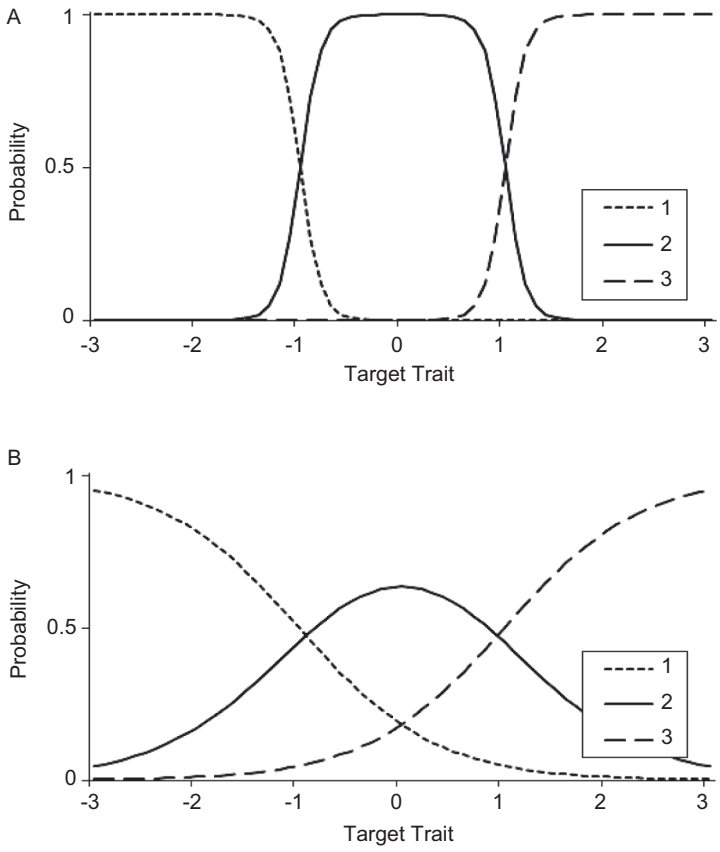

FIGURE 7.2. Response functions for a polytomous item with high (A) and low (B) discrimination.

differentiating, or discriminating, between individuals having different levels of a target trait; if one knows how someone scored on this item, one has rich information about the respondent's level of target trait.

In practice, the level of discrimination depicted in Figure 7.2A is rarely observed, and the assessment developer must be content with a more modest discrimination whereby the response to the item does not provide completely unambiguous information concerning the range of target trait in which the respondent is located. An example of such an item is presented in Figure 7.2B. For this item, most regions of the target trait continuum are associated with multiple response categories having a substantial chance of occurring, and thus the response categories are not highly effective at differentiating between individuals of different target trait levels. That is, for most target trait levels substantial overlap exists between two or more response categories, and thus the response category does not provide unambiguous information concerning the target trait value of the respondent. The overlap reflects measurement error in the obtained responses to the item; a given respondent may not provide the same response to the same item on two independent occasions, all other things being equal. Items with unacceptably low levels of discrimination have high overlap between the response categories, such that there is little correspondence between the target trait and the response categories; each category is just about equally likely to be observed regardless of target trait. Items with an absence of discrimination are those for which the chance of selecting a given response category is constant across all target trait levels, and thus score level provides no information about trait level.

Low item discrimination typically results from one of two situations. The first situation is the presence of an ambiguity in the item (e.g., item stem or response options), such that different individuals interpret the item content in different ways, which

introduces high levels of error into responses to the item. The second situation is item content that is not aligned with the intended target trait. In this instance, the item content may be clear and unambiguous, but the response categories are not effectively grouping respondents with respect to the intended target trait but rather with respect to some other trait.

A widely used item format in assessing cognitive abilities and skills is the multiple-choice format, whereby the respondent is instructed to select one of several (usually between three and five) options, of which one is the correct option and the others are distractor options, or simply distractors. These items are commonly scored dichotomously as correct and incorrect. A highly discriminating multiple-choice item demonstrates a tight correspondence between item response (correct or incorrect) and target trait level; individuals selecting the correct response tend to lay in a specific portion of the target trait continuum, and individuals selecting an incorrect response tend to lay in a lower region of the target trait continuum. As an example, Figure 7.3A displays the IRFs for a highly discriminating multiple-choice item. Note that a correct response indicates that the respondent's target trait level is highly likely to exceed 0.2, and an incorrect response indicates that the respondent's target trait level is highly likely to be below −0.2. Thus, the response to this item provides a large amount of information concerning the respondent's level of target trait. In contrast, Figure 7.3B presents the response functions for a multiple-choice item having a much more modest level of discrimination; there is a much greater overlap of the response functions, such that for much of the target trait continuum there is a substantial chance of

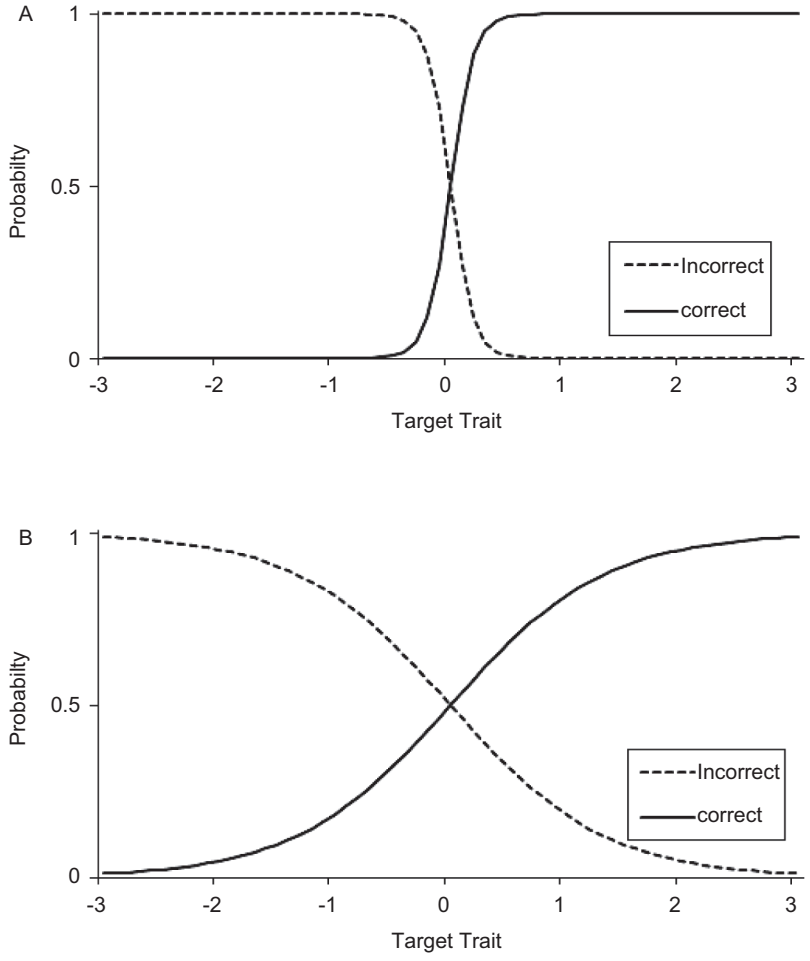

FIGURE 7.3. Response functions for a dichotomous item with high (A) and low (B) discrimination.

observing either a correct or an incorrect response. In this instance, the link between item response and level of target trait is weaker, and thus the item provides less information concerning the respondent's level of target trait.

Item Difficulty

Using the IRF framework described earlier, one can also more formally define item difficulty. The difficulty of an item is determined by the location of the target trait continuum differentiated by the item's response categories (i.e., the level of target trait at which the chance of different response categories occurring changes dramatically). If the response categories of an item tend to differentiate between individuals of relatively high target trait levels, then the item is relatively high in difficulty. In contrast, if the response categories of an item tend to differentiate between individuals of relatively low target trait levels, then the item is relatively low in difficulty (often referred to as *easy*). The term *difficulty* derives its name from testing applications whereby items that differentiate between moderate-ability and high-ability individuals are referred to as *difficult,* and items that differentiate between low-ability and moderate-ability individuals are referred to as *easy*. In this instance, a difficult item is one for which a correct response is only likely to occur for individuals with high levels of the target trait.

As an example of item difficulty, consider the IRFs for two different items displayed in Figure 7.4. Both items' response functions have the same shape and thus share an identical discrimination. However, the first item (A) differentiates between individuals with low levels of target trait and thus has low difficulty. For this item, any individual with a moderate or high level of target trait is expected to select the same item response category (3), and thus this item provides negligible information to differentiate between individuals at high target trait levels. Rather, the item's response categories serve to differentiate between individuals having very low, moderately low, and slightly low levels of target trait. In contrast, the second item (B) is high in difficulty. The response categories of this item differentiate among individuals with slightly high, moderately high, and very high levels of the target trait. Because

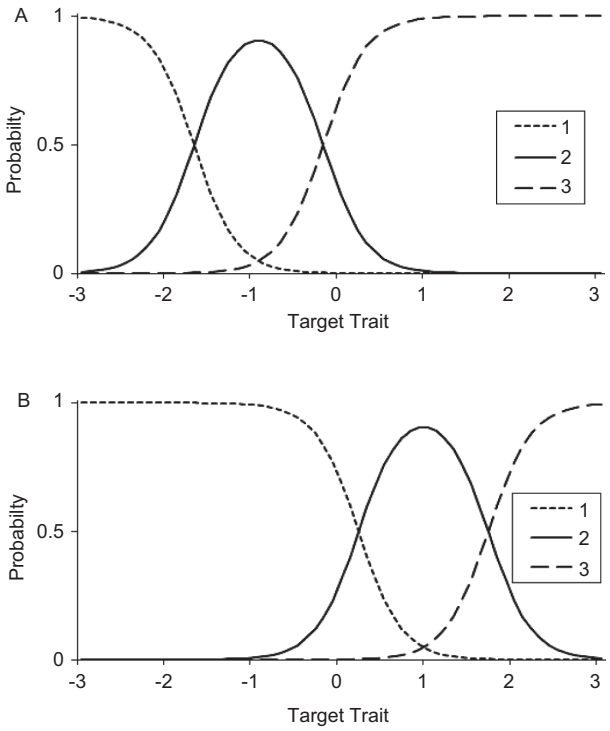

FIGURE 7.4. Response functions for a polytomous item with low (A) and high (B) difficulty.

all individuals who are low in the target trait are expected to receive the same response (1), this item provides negligible information for individuals who are low in target trait.

The concept of item difficulty is important because assessments are developed with the intent of providing a particular level of information across the target trait continuum; some assessments are intended to provide a relatively constant level of information across the entire continuum, and other assessments are developed with the intent of providing a high level of information for only a particular range of target trait levels, as would be the case for assessments used to categorize individuals at critical points along the target trait continuum. The difficulty of the items contained in the assessment determines whether the assessment provides information at target trait levels in a manner that meets the intent of the assessment. As a result, the appropriateness of any given item's difficulty must be considered in the context of the assessment's purpose. Clearly, any item having a difficulty so extreme as to not differentiate between any respondents in the intended population is not a useful item. However,

it may be important to include numerous items of high difficulty (i.e., if the assessment is being used to differentiate between individuals at the high end of the target trait continuum) or low difficulty (i.e., if the assessment is being used to differentiate between individuals at the low end of the target trait continuum) or a range of low, medium, and high difficulty (i.e., if the assessment is being used to differentiate between individuals at all levels of the target trait).

Category Use

A factor that plays an important role in the discrimination of an item is the extent to which the categories of the item are being selected in a meaningful way that allows the responses to differentiate between individuals of different trait levels in a systematic fashion. It is thus often useful to examine how each response category of an item is being used. Any category with negligible use (e.g., a category having only 1% of responses) should be attended to because this category is providing negligible information for differentiating between individuals of different target trait levels. For example, if the lowest two categories of a rating scale item are not being selected by the respondents, then the assessment developer should consider revising the item so that the response categories can be used more effectively to differentiate between individuals; the lack of use of the lowest two categories is compromising the potential information generated by the item.

For multiple-choice items, the use of response options can yield useful information concerning potential problems with the item. Distractors that have a very low rate of selection (say, less than 5%) or a very high rate of selection (say, more than 70%) may have an undesirably low or high level of attraction for the respondents and should be considered for revision. A distractor with a very high selection rate may contain a deceptive element that is inappropriately tricking respondents into selecting the option. In addition, a correct option that has a selection rate substantially below that expected by chance alone (i.e., for a four-option item, the correct option should have a selection rate of 25% by random guessing alone) indicates a potential deceptive property in the item causing the correct option to be overly unattractive to the respondents. In this instance, the assessment developer is advised to review the content of the item for potential problems.

OBSERVED SCORE APPROACHES FOR ESTIMATING ITEM PROPERTIES

Incipient procedures for quantifying item difficulty and discrimination were founded in relatively simple statistical procedures computed using nothing more than the raw item scores and their associated sums. This set of procedures was developed within the classical test theory framework (Crocker & Algina, 1986), in which the level of target trait is approximated using the observed summated score (i.e., the sum across the items of the assessment), thus the name *observed score approaches*. These approaches have the advantage of being computationally simple, requiring relatively small sample sizes for adequate stability of the discrimination and difficulty estimates, and can be obtained using widely available statistical software (e.g., SPSS).

Item Discrimination

Because the assumption is that the score levels are monotonically increasing (i.e., successively higher score levels become more likely to be attained as the level of target trait increases), item discrimination can be evaluated through consideration of how much the mean (or expected) item score increases as the target trait increases. A highly discriminating item is one for which the mean item score is substantially higher for individuals with high target trait levels than for individuals with low target trait levels, and a nondiscriminating item is one for which the mean item score is the same for individuals with low and high target trait levels. A suitable approach for measuring the magnitude of discrimination is the correlation between the item response and the total score used to measure the target trait, where the total score is typically given by the sum of the item-level scores (or a weighted sum). This correlation is commonly referred to as the *item–total correlation* (also known as the *point-biserial correlation*). A value of zero indicates no discrimination, and a value of 1 indicates perfect discrimination.

A common modification to the item–total correlation is to adjust the total score used in computing it for each item such that the adjusted score is the sum of all items other than the item in question (i.e., the total score obtained with the omission of the item for which the item–total correlation is being computed). The adjusted item–total correlation is referred to as the *corrected item–total correlation*. The rationale for this adjustment is to avoid inflated measures of discrimination caused by the dependence of the total score on the item under investigation. Like the item–total correlation, the corrected item–total correlation should assume values between 0 (no discrimination) and 1 (perfect discrimination).

Acceptable values of the item–total correlation vary depending on the form of the item, the breadth of content domain underlying the assessment, and the reliability of the total score. Nonetheless, assessment developers typically make use of general guidelines of acceptable values to facilitate interpretation of the information provided by an item and guide the associated decisions concerning item revision and removal. Items with an item–total correlation (or corrected item–total correlation) near zero (say less than .1 in magnitude) are providing virtually no information about the respondent's target trait level and can thus be removed from the assessment without any loss of overall information. Items with an item–total correlation value on the order of .1 to .3 are providing a relatively small amount of information and should be flagged for removal or revision. Items with item–total correlation values on the order of .3 to .5 are providing a moderate amount of information, and values greater than .5 reflect large amounts of information. The magnitude of the item–total correlation will often be lower for multiple-choice items than for polytomously scored items (i.e., rating scale items with three or more score levels). As a result, an item–total correlation value of .2 or greater is often viewed as acceptable for multiple-choice items, and values exceeding .4 are often viewed as quite good. In addition, small item–total correlations are only acceptable when the content domain is broad (e.g., verbal ability), such that no single item defines the target trait. In situations in which the content domain is very narrow (e.g., self-efficacy of third-grade reading comprehension), substantial overlap will exist in the content of the assessment's items, and thus small item–total correlations are typically viewed as unacceptable.

Although it would violate the assumption of monotonicity, negative item–total correlations are occasionally observed in practice, and such a situation requires attention from the assessment developer. Negative discriminations have two potential causes. The first potential cause is just a poorly constructed item. The second potential cause, which is often the root of large negative item–total correlations, is the failure to appropriately code the score levels of items such that the item score levels are monotonically increasing with target trait level. With multiple-choice items, this reflects an incorrect specification of the correct option, and with rating scale items, this often occurs when reverse-worded items (often referred to as *negatively worded items*) are not appropriately reverse coded.

Although having items with high item–total correlations usually represents a desirable scenario, at a point having all (or most) item–total correlations being too high becomes an undesirable situation. When all (or most) items have an item–total correlation that is extremely high (say, more than .8), then the items are highly intercorrelated and are thus providing redundant (or overlapping) information. If all items are providing highly redundant information, then there is no need to include all items on the assessment; multiple items provide no more information than a single item. Furthermore, the presence of items providing highly redundant information indicates that the universe of content underlying the assessment has likely not been adequately sampled, which can compromise the validity of the scores generated by the assessment. As an extreme example of this situation, consider an assessment containing 12 identical items. Presumably, each respondent would provide the same response to each of the 12 items, each pair of items would be perfectly correlated with one another, and the resulting item–total correlations would be extremely high (near 1.0) for all items. However, nothing is gained from administering the same item 12 times; just one of the items would afford the same information as all 12 items. To make matters worse, the

content specifications underlying the assessment are unlikely to be adequately represented with a single item, and thus the validity of the obtained scores would be called into question.

Let us apply these guidelines to an example of a 10-item rating scale used in the assessment of anxiety, for which each item has response categories 1, 2, 3, and 4. Table 7.1 presents the results of an item analysis for these 10 items on the basis of a sample of 2,000 respondents. The first column of results presents the corrected item–total correlation. These values range from .08 (Item 3) to .68 (Item 7). Two items display notably low levels of discrimination: Item 3 (.08) and Item 6 (.17). For these items, the expected item score is not adequately predictive of the target trait. These items should be reviewed for ambiguous content, or content that is not aligned with the target trait for this scale, and revised accordingly or removed altogether. It is relevant to note that the reliability (estimated using Cronbach's alpha) of the total summated scale scores is equal to .73 when all 10 items are included in the assessment but increases to .80 after removing Items 3 and 6. Thus, removing the two items with poor discrimination served to increase the reliability of the scores generated by the instrument.

Information concerning the potential causes of low discrimination can be furnished through examination of the response distribution across the four score levels of the items on the rating scale. The two items demonstrating low discrimination (Items 3 and 6) both had very low use of the middle two response categories, such that more than 90% of the responses to Item 3 reside in the two extreme response categories (1 and 4) and more than 80% of the responses to Item 6 reside in the two extreme response categories. For these two items, respondents are not effectively using the full range of response categories, which may be contributing to the low discrimination of the items. An attempt to revise these items may be well served by considering why the two middle score levels are not being used in responding to these items.

As a second example, consider a 26-item multiple-choice assessment of analytic skills administered to 5,000 respondents for which each item contained five response options: the correct option and four distractor options. Each item was coded dichotomously as correct (i.e., the correct option was selected) or incorrect (one of the four distractors was selected). The leftmost portion of Table 7.2 displays the results concerning item discrimination under the observed score framework. Two items stand out as having an unacceptably low corrected item–total correlation; namely, Items 4 and 13 (.07 and −.05), respectively. The content of these items should be reviewed, and the items should be revised accordingly or removed from the assessment. Two items have a corrected

TABLE 7.1

Item Properties for a 10-Item Scale

	Observed score approach		Item response theory approach				
Item	Corrected item–total correlation	Mean item response	a	b_1	b_2	b_3	b_+
1	.37	2.39	0.93	–0.76	0.17	1.20	0.20
2	.40	3.17	1.14	–2.34	–0.99	–0.26	–1.20
3	.08	2.62	0.18	–1.73	–1.00	0.02	–0.90
4	.49	2.63	1.40	–1.07	–0.30	0.76	–0.20
5	.53	1.47	2.55	0.70	1.20	2.32	1.41
6	.17	2.38	0.37	–0.48	0.36	1.46	0.45
7	.68	2.46	3.06	–1.06	0.03	1.20	0.06
8	.46	3.15	1.35	–1.97	–1.18	0.01	–1.05
9	.42	1.31	1.91	1.12	1.99	2.65	1.92
10	.61	2.78	2.63	–0.92	–0.81	0.80	–0.31

TABLE 7.2

Item Properties for a 26-Item Multiple-Choice Assessment

	Observed score approach		Item response theory approach	
Item	Corrected item–total correlation	Item mean	*a*	*b*
1	.42	0.54	1.19	–0.16
2	.41	0.83	1.89	–1.25
3	.33	0.43	0.85	0.38
4	.07	0.23	0.19	6.59
5	.45	0.55	1.33	–0.20
6	.41	0.45	1.13	0.24
7	.42	0.54	1.16	–0.15
8	.47	0.61	1.47	–0.44
9	.44	0.80	2.01	–1.11
10	.48	0.42	1.43	0.30
11	.34	0.69	0.99	–0.94
12	.43	0.65	1.33	–0.62
13	–.05	0.11	0.15	13.91
14	.44	0.77	1.83	–1.01
15	.24	0.83	0.83	–2.11
16	.36	0.62	0.99	–0.59
17	.37	0.50	0.95	–0.01
18	.42	0.49	1.15	0.04
19	.35	0.41	0.92	0.46
20	.40	0.83	1.77	–1.30
21	.23	0.77	0.69	–1.88
22	.34	0.63	0.92	–0.67
23	.37	0.49	0.96	0.05
24	.39	0.49	1.04	0.03
25	.34	0.51	0.86	–0.02
26	.51	0.59	1.66	–0.35

item–total correlation coefficient that is marginal in magnitude: Item 15 (.24) and Item 21 (.23). These two items may be reviewed for problematic content (recall, however, that multiple-choice items will often have an item–total correlation in the .20 range). The other items of the assessment demonstrate item–total correlations between .33 and .51 and thus demonstrate acceptable levels of discrimination.

Item Difficulty

Under the observed score approach, item difficulty is measured using the mean of the responses to the item. That is, for each item of the assessment, one obtains the mean value across the responses to the item, and these means reflect the relative difficulty of the items. The lower the mean, the more difficult the item was for the sample of respondents. Notice the inverse relationship that exists between the mean and difficulty; lower means imply higher difficulty. In the case of items that are dichotomously scored as 0 and 1 (e.g., incorrect and correct), the mean is equivalent to the proportion of individuals scoring 1 (i.e., the proportion correct). For this reason, the mean of 0/1 scored items is commonly referred to as the *p* value (*p* for proportion) for the item. Again, note that lower *p* values reflect higher difficulty.

The means for each item of the 10-item anxiety assessment are given in the third column of Table 7.1. The items on this scale had score levels of 1, 2, 3, and 4, and thus mean values could range from 1 (all respondents selected a 1) to 4 (all respondents selected a 4). Across the 10 items of the scale, one can see that two items were of relatively low difficulty (Items 2 and 8), six items were of moderate difficulty (Items 1, 3, 4, 6, 7, and 10), and two items were of high difficulty (Items 5 and 9). Because item difficulty reflects the level of target trait about which the item provides information, one can infer that this group of 10 items provides information across a wide range of target trait levels: (a) Items 2 and 8 (low difficulty) provide information differentiating between individuals with low levels of the target trait; (b) Items 1, 3, 4, 6, 7, and 10 (moderate difficulty) provide information differentiating between individuals with moderate levels of the target trait; and (c) Items 5 and 9 (high difficulty) provide information differentiating between individuals with high levels of the target trait.

An important consideration when examining item difficulty is to identify items having a particularly extreme level of difficulty, because these items may be providing information for levels of the target trait that are more extreme than those observed in the population under investigation. For the 10-item anxiety assessment, any item with a mean that is very close to 1 or 4 should be flagged for review. One can see that one item meets this criterion: Item 9 has a mean of 1.31, indicating that it provides

information about individuals who are extremely high in the target trait. For this item, the vast majority of respondents (80%) scored a 1, and only 8% scored a 3 or 4. As a result, this item serves to discriminate between only the upper 20% of the respondents; for the lower 80% of respondents this item provides no differentiating information (because all individuals have the same score of 1 on this item). The assessment developer would need to decide whether Item 9 was of appropriate difficulty for the population of interest.

As a second example of item difficulty, consider the results for the multiple-choice assessment of analytic skills (Table 7.2). Because these items were scored dichotomously (0 = *incorrect,* 1 = *correct*), the means can range from 0 (nobody selected the correct option) to 1 (everyone selected the correct option). The means span a wide range (0.11 to 0.83), indicating that the assessment as a whole is providing information across a wide range of trait levels. There are, however, several items that display a difficulty level extreme enough to warrant review. Item 13 is the most extreme with respect to difficulty, having a mean of 0.11 (only 11% of individuals selected the correct response). Note that this rate of correct response is below what one would expect by random guessing alone, which provides evidence that one or more of the distractors is unusually attractive for this sample. Examining the response distribution across the options, one finds that Option C was selected by 64% of the sample, suggesting that this distractor should be given particular attention in the item review process. One other item, Item 4, demonstrated a low mean value (.23), which could warrant additional content review. Four items (Item 2, 9, 15, and 20) had a mean value of .80 or greater and thus were extremely low in difficulty for this sample. This may be more items than necessary at this low level of difficulty, and the assessment developer may consider replacing one or more of these items with an item having a more moderate level of difficulty. In making this decision, however, the assessment developer would need to consider the level of information desired across the target trait continuum; if having a high level of information at low levels of target trait is important, then an argument can be made for retaining all of the items with low difficulty.

Sample Size Requirements

A question that often arises in planning pilot testing and conducting item analyses is what constitutes an appropriate sample size for computing the observed score estimates of item discrimination and difficulty. There is no single answer that applies to all settings, but some rough guidelines can be considered. As a very general rule of thumb, a sample size of 200 respondents will typically provide an adequate level of stability for both the item mean and the item–total correlation. A sample size of 100 can be viewed as a lower threshold of acceptability; although this sample size is expected to provide adequate stability for the item mean to make a statement concerning item difficulty, the item–total correlation may not have adequate stability to appropriately differentiate among low, moderate, and high discrimination. In instances in which the sample size is notably less than 100 (say, 50–70), the results of an item analysis using observed score methods should be interpreted with extreme caution (particularly the item–total correlation) and should be used only to inform general trends in difficulty and discrimination and extreme item properties.

ITEM RESPONSE THEORY APPROACHES FOR ESTIMATING ITEM PROPERTIES

Although an item analysis conducted using the observed score approach provides useful information concerning item properties, item response theory (IRT) provides a flexible model-based approach that has numerous desirable properties for quantifying item properties, evaluating item quality, developing assessments, and evaluating the properties of scores generated by assessments (de Ayala, 2009; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Wright & Masters, 1982). The general idea behind IRT is to generate the IRFs for each item (such as those presented in Figures 7.2, 7.3, and 7.4) using a mathematical equation that contains one or more parameters related to the item's discrimination and difficulty. The parameters of the equation are estimated from the responses made to the item, and the obtained parameter estimates are then used to interpret the item's discrimination and difficulty.

This section describes how item discrimination and difficulty are measured using IRT. I should point out up front that this description is not intended to provide a comprehensive presentation of IRT itself; readers interested in a broader presentation of IRT are referred to Chapter 6 in this volume and other texts that provide broader discussions of the IRT framework and applications (de Ayala, 2009; Embretson & Reise, 2000; Hambleton et al., 1991).

Dichotomous Items

For dichotomous items, there are only two score levels (usually coded as 0 and 1) for the item, most commonly corresponding to incorrect (0) and correct (1) scores obtained from multiple-choice or short-answer item formats. In IRT, a mathematical equation is used to specify the probability of each score level as a function of the target trait (see the example in Figure 7.3). A widely used equation for dichotomous items is the two-parameter logistic model, which specifies the probability that the item response (Y) is scored as 1 (e.g., correct response) at a particular level of target trait (θ) using

$$P(Y=1\,|\,\theta=\frac{\exp\left[1.7a(\theta-b)\right]}{1+\exp\left[1.7a(\theta-b)\right]}. \qquad (7.1)$$

The specific form of this model is of little concern in this chapter (more discussion of this is presented in Chapter 6, this volume). However, what is of immediate relevance to this chapter are the two item parameters contained in the model: a and b. The discrimination of the item is determined by the a parameter, and the difficulty of the item is determined by the b parameter. Furthermore, a, b, and θ are all on a common metric, which allows the values of the a parameter and b parameter to be interpreted in relation to the target trait continuum and affords particularly useful interpretations of the information provided by each item. A detailed description of the a parameter and b parameter is provided next.

Item discrimination. The a parameter reflects item discrimination because it determines how steep the IRFs are, or how dramatically the probability of observing a 1 shifts from very low (near zero) to very high (near 1). As the a parameter increases in magnitude, the IRF increases in its steepness, and discrimination increases. The lowest possible magnitude of the a parameter is zero, which corresponds to the absence of discrimination. When $a = 0$, the IRF for a score level of 1 is a horizontal line, such that the probability of a 1 is constant for all levels of target trait (and thus the probability of a 0 is also constant for all target trait levels). In this situation, knowing an individual's response to the item provides no information concerning the individual's target trait level. In contrast, as the a parameter approaches infinity, the IRF approaches a vertical form at its midpoint (Figure 7.3A approaches this situation, having $a = 10$), and the score levels of 0 and 1 are associated with distinct ranges of the target trait continuum. An illustration of how the a parameter changes the steepness of the IRF is shown in Figure 7.3. The item presented in Figure 7.3A has a higher a parameter ($a = 10$) than the item presented in Figure 7.3B ($a = 1.5$).

A negative value of the a parameter ($a < 0$) reflects a violation of the monotonicity condition in the current coding of the item, which can result from either an incorrect coding of the correct option or a substantial problem with the item's content. In this situation, the score level of 1 becomes less likely to occur as the target trait level increases, thus reflecting an altogether undesirable situation. This is the IRT equivalent of having a negative item–total correlation coefficient (within the observed score approach). Any item having a less than 0 should be checked for the appropriate coding of the correct option, and if no coding error can be found, then the item should be revised or removed from the assessment.

In practice, using a standard IRT metric, a parameter values tend to range between 0 and 2 (although values exceeding 2 are occasionally observed). Values near 0 reflect very poor (or no) discrimination, and any item having a less than 0.4 should be considered for revision or removal from the assessment. Such items provide little information concerning the respondent's target trait level and thus contribute little to the precision of the target trait estimate. Items having an a parameter between 0.4 and 0.8 are demonstrating a moderate

level of discrimination; these items may not require revision or removal but may be considered for review to determine whether there is any way to improve on the items' discrimination. Such items provide a meaningful amount of information concerning the respondent's target trait level and are thus making a useful contribution to the precision of the target trait estimate. Items with an *a*-parameter value greater than 0.8 have adequate discrimination, and items with an *a*-parameter value in excess of 1.5 have very strong discrimination.

As an example of using the *a* parameters to quantify item discrimination in dichotomously scored items, Table 7.2 presents the *a*-parameter values for the items of the multiple-choice assessment of analytic skills. Notice that two items have unacceptably low levels of discrimination: Item 4 ($a = 0.19$) and Item 13 ($a = 0.15$). These two items should be flagged for review and potential revision or removal. All other items have acceptable levels of discrimination, with several items (Items 2, 9, 14, 20, 26) demonstrating very high discrimination. It is relevant to note (from Table 7.2) that the magnitude of discrimination indexed by the *a* parameter yields similar information to that provided by the corrected item–total correlation; the two items with very low *a*-parameter values (Items 4 and 13) also demonstrated near-zero corrected item–total correlations (observed score approach), and the items with the highest *a*-parameter values (Items 2, 9, 14, 20, 26) also demonstrated relatively high corrected item–total correlations. This finding is expected because under the specific condition of a normally distributed target trait, the *a* parameter is a simple transformation of the correlation between the item and the target trait (Lord & Novick, 1968). Note, however, that the relative ordering of the *a*-parameter values in Table 7.2 differs slightly from that of the corrected item–total correlation, a result that is often observed in practice. Thus, although the two approaches will generally provide consistent results concerning discrimination, they will typically not yield identical results.

Item difficulty. The *b* parameter reflects item difficulty by specifying the target trait level at which the information provided by the item is highest. Recall that the *b* parameter is on the same metric as the target trait (θ), and thus the value of the *b* parameter can be interpreted with respect to specific target trait values. Using the two-parameter logistic model, the *b* parameter reflects the target trait value at which the IRFs intersect (the target trait value at which the probability of a 0 and a 1 are both equal to .5). In Figures 7.3A and 7.3B, the IRFs intersect at a target trait value of 0, and thus $b = 0$; these items provide the maximum information for individuals having a target trait level of 0. As the *b* parameter increases, the item difficulty increases, and the item provides information about individuals with a higher level of the target trait. An advantage of interpreting item difficulty in the IRT framework is that the item information can be interpreted directly in relation to the target trait of respondents. That is, if *b* equals -1, then one knows that this item provides information for individuals at a target trait level of -1. As a result, examining the distribution of *b*-parameter values provides quick and rich information concerning where (i.e., for which target trait values) the assessment is providing information. Naturally, this should also be coupled with consideration of item discrimination (*a* parameters) to inform the issue of how much information. The observed score approach for measuring item difficulty (i.e., the mean item response) has no implicit way of accomplishing this; although the mean item response informs the relative difficulty of the items, it does not directly address which target trait values are being informed by each item.

In general, assuming the standard IRT metric, the *b*-parameter value will typically range between -3 and 3, and whether the difficulty is too extreme should be determined by where the item difficulty resides in relation to the distribution of target trait values observed in the population for which the assessment is intended. An item with a *b* parameter that is not aligned with target trait values contained in the population of interest is not providing useful information for the population of interest (i.e., the item is too easy or too difficult for the population of interest). The particular distribution of *b* parameters desired for items of an assessment will depend on the intended use of the assessment and the population of interest, but in general items with *b* parameters

that fall outside the range of the target trait representative of the population of interest should be considered for revision or removal.

As an example of how to interpret the *b* parameter, consider again the multiple-choice assessment of analytic skills, for which the relevant *b*-parameter estimates are shown in Table 7.2. Two items (Items 4 and 13) demonstrated extreme *b*-parameter values of 6.59 and 13.91. These items are clearly too difficult to provide useful information on the population of interest. These same items also demonstrated unacceptably low values of the *a* parameter and are thus items with substantial problems (note that the joint estimation of *a* and *b* parameters leads to a confounding of the obtained estimates, such that a low *a*-parameter estimate will often be associated with an extreme *b*-parameter estimate). Several items (Items 15 and 21) demonstrated very low *b*-parameter values and thus provide information about respondents with very low target trait values (these same items had relatively high mean item responses and thus low item difficulty values under the observed score approach).

Polytomous Items

The interpretation of discrimination and difficulty for polytomous items (i.e., items with more than two score levels) extends from the discussion of dichotomous items. However, because polytomous items contain more than two score levels, the process of quantifying item difficulty and discrimination is more complicated than that for dichotomous items. Although several IRT models are commonly used for polytomous items, I focus on one model known as the graded response model (GRM) because of its widespread use in the literature. Although the discussion focuses on the GRM, the concepts and interpretations presented here generalize to other polytomous models.

To describe how discrimination and difficulty are defined using the GRM, consider a polytomous item with J score levels, such that a response to the item is denoted by $Y = 1, 2, \ldots, J$. For example, an item having three score levels would have $J = 3$, such that the response to the item could assume the values $Y = 1, 2, 3$ (as is the case for the items depicted in Figures 7.2 and 7.4). For an item having J score levels, the GRM specifies the IRFs (just as with those in Figures 7.2 and 7.4) using a total of J parameters. The first of the J parameters is an *a* parameter that is similar to the one used for the two-parameter model for dichotomous items (Equation 7.1). Not surprisingly, this *a* parameter serves as a measure of discrimination, more details of which are provided next. The remaining $J - 1$ parameters are a series of difficulty parameters, and these are similar to the difficulty parameter (*b*) used for the two-parameter logistic for dichotomous items (Equation 7.1), with the exception that there is more than one such difficulty parameter. For an item having three score levels ($J = 3$), there will be $J - 1 = 2$ such *b* parameters, denoted as b_1 and b_2. For an item with four score levels ($J = 4$), there will be $J - 1 = 3$ such *b* parameters denoted as b_1, b_2, and b_3, and so forth. The values of the *b* parameters determine where the IRFs are located along the target trait continuum, more details of which are provided next. In the IRT literature, the *b* parameters of the GRM are commonly referred to as *location* or *transition parameters* because they reflect the relative difficulty of transitioning to successively higher score levels (described in more detail later). For simplicity, I refer to these parameters simply as *b* parameters in the ensuing discussion.

Item discrimination. Discrimination in polytomous items can be quantified by the value of the *a* parameter in the GRM. As the *a* parameter increases in magnitude, the IRFs become steeper, and thus each score level corresponds more tightly to a particular range of the target trait continuum. The magnitude of the *a* parameter for the GRM is analogous in interpretation to dichotomous items; it can range from 0 to infinity (negative values correspond to a violation of monotonicity, values near zero reflect IRFs that are very flat and overlapping). As an example, consider the two polytomous items depicted in Figure 7.2. The item in Figure 7.2A has $a = 10$, which is associated with a high level of discrimination because each score level is highly associated with distinct ranges of the target trait continuum (note the very steep IRFs for this item). This level of discrimination is substantially higher than what would typically be observed in practice. In contrast,

the item in Figure 7.2B has $a = 1.5$, which is associated with a lower degree of discrimination (although in practice this level of discrimination would be quite acceptable) and thus relatively flatter IRFs. The values of the a parameter for polytomous items can be interpreted using similar criteria to those outlined for dichotomous items. Values less than 0.4 are relatively low and warrant further review, values greater than 0.8 are satisfactory, and values of 1.5 or higher are considered high.

As an example, consider the 10-item anxiety assessment for which the item discrimination and difficulty parameters are presented in the right side of Table 7.1. Item 3 has a very low discrimination value ($a = 0.18$), and Item 6 also has a notably low discrimination value ($a = 0.37$). These items should be subject to a content review and potentially revised or removed from the scale. Not surprisingly, these were the same items flagged as having low discrimination using the observed score approach (corrected item–total correlations of .08 and .17). All other items demonstrate good discrimination, with several items (Items 5, 7, and 10) demonstrating very high discrimination.

Item difficulty. Difficulty of polytomous items modeled using the GRM is quantified through the b parameters. As described earlier, a polytomous item having J score levels will have $J - 1$ b parameters, and each b parameter corresponds to a particular score level. The value of b_1 equals the level of target trait required to have a probability of .5 of obtaining any score greater than 1, the value of b_2 equals the level of target trait required to have a probability of .5 of obtaining any score greater than 2, the value of b_3 equals the level of target trait required to have probability of .5 of obtaining any score greater than 3, and so on. Thus, each successively higher b parameter represents the target-trait-required score in a successively higher range of item score levels. As a result, the b parameters should always be increasing in value ($b_1 < b_2 < b_3$, etc.) to reflect the fact that higher score levels are associated with higher levels of target trait (which is the monotonicity condition). Note that each b parameter reflects the difficulty of transitioning to successively higher score levels, and thus the values of the b parameters can be used to evaluate the difficulty of each transition to higher score levels.

As an example of how to interpret item difficulty under the GRM, consider the items shown in Figure 7.2 for which there are three score levels and thus only two b parameters (b_1 and b_2). Although the two items shown in Figure 7.2 have different a parameters, they have identical b parameters: $b_1 = -1.0$ and $b_2 = 1.0$. The value of b_1 indicates that a target trait value of -1.0 is required to have a probability of .5 of scoring a 2 or 3 on the item (and thus also a probability of .5 of scoring a 1), and thus the first transition (from a 1 to anything higher than a 1) provides information about the target trait for individuals in the general area of -1.0. Similarly, the value of $b_2 = 1.0$ indicates that a target trait value of 1.0 is required to have a probability of .5 of scoring a 3 on the item (and thus also a probability of .5 of scoring a 1 or 2), and thus the second transition (from a 1 and a 2 to a 3) provides information about the target trait for respondents in the general area of 1.0. Note that the b parameters are increasing in value—any violation of this property would be an immediate sign that monotonicity is not maintained (individuals with a higher target trait level are obtaining lower score levels than individuals with a lower target trait level). Taking the values of b_1 and b_2 together indicates that the item provides most of its information primarily between target trait levels of -1.0 to 1.0 and thus generates information about respondents having moderate target trait levels. Notice that because polytomous items provide information across two or more transitions (located at target trait values of -1.0 and 1.0 for this item), they tend to provide information over a wider target trait range than individual dichotomously scored (e.g., multiple-choice) items.

Consideration of the individual b parameters provides rich insight into the range of target trait about which the item provides information, insight that is simply not available in the observed score approach in which the mean item response is taken as a measure of item difficulty. Despite this advantage, having a single overall item-level index of difficulty is often still useful so that one can make general statements concerning the relative difficulty of the items, that is, which items provide information about low,

moderate, and high levels of target trait. Such an index of overall item difficulty can be readily obtained by calculating the average *b*-parameter value (denoted here as $b_{+)}$. For example, if a polytomous item having four score levels has *b*-parameter values of $b_1 = -2.2$, $b_2 = -1.6$, and $b_3 = 0.2$, then $b_+ = -1.2$, suggesting that the item provides information about a range of target trait that is centered approximately on −1.2 and would thus typically be viewed as a relatively low-difficulty item. The actual characteristics of the range of target trait for which the item provides information would be informed by the individual *b* parameters; because two of the *b* parameters are below −1.2, slightly more information is provided below −1.2 than above −1.2. In the case of the items presented in Figure 7.2 (for which $b_1 = -1.0$ and $b_2 = 1.0$), the overall item difficulty can be summarized by $b_+ = 0$. Similarly, for the items presented in Figure 7.4 ($b_1 = -1.7$ and $b_2 = -0.2$ for Figure 7.4A, and $b_1 = 0.2$ and $b_2 = 1.7$ for Figure 7.4B), the overall item difficulty can be summarized by $b_+ = -0.95$ (Figure 7.4A) and $b_+ = 0.95$ (Figure 7.4B).

The pattern of the *b* parameters informs not only the range of target trait about which the item provides information, but also the extent to which adjacent score levels provide unique information. If two adjacent *b* parameters are very close to one another (say, within 0.3 units of one another), then the distinction between the successive response categories is negligible—that is, the successive categories are not providing information about unique ranges of the target trait continuum (unless the discrimination is very high). If there are wide gaps between *b* parameters (say, more than 2 units), then the item likely provides information at localized regions of the target trait continuum. Similarly, if there is an extreme *b* parameter, then the evidence shows that the associated extreme score level is not functioning as desired.

Let us now apply these concepts to the 10-item assessment of anxiety, for which the *b* parameters are presented in the right side of Table 7.1. These items contain four score levels, and thus there are three *b* parameters (b_1, b_2, b_3) and an overall index of difficulty, b_+. Item 9 is the most difficult item ($b_+ = 1.92$), having information that is centered approximately at a target trait value of 1.92, and allocates its information at the target trait range of 1.12 (the value of b_1) to 2.65 (the value of b_3). This item provides substantial information about individuals very high in the target trait but provides little information about individuals having a target trait value that is moderate or low. In contrast, Item 2 is the easiest, having its information centered at approximately −1.20 ($b_+ = -1.20$) and providing its information for individuals on the target trait range of approximately −2.34 and −0.26. Items 1, 4, and 7 provide information across a target trait range centered on approximately 0 and thus allocate their information to respondents having moderate target trait levels in the approximate range of −1 to 1.

Item 10 of Table 7.1 provides an interesting case, in which the values of b_1 and b_2 are very close ($b_1 = -0.92$ and $b_2 = -0.81$). This close proximity indicates that the transition from a 1 to a 2 and the transition from a 2 to a 3 provide largely redundant information; that is, there is no meaningful information generated by the distinction between a 2 and a 3. As a result, the assessment developer should review this item and the anchors assigned to the score levels in an attempt to identify why these two score levels are not differentiating between individuals. All other items of the scale have adequate separation between the *b* parameters, suggesting that each transition provides unique information useful in differentiating between individuals of different target trait levels.

Sample Size Requirements

Despite the advantageous properties of the IRT framework, its primary disadvantage compared with observed score methods is the larger sample sizes required to obtain adequately stable parameter estimates. The required sample size for stable parameter estimation is a complicated issue because it depends on the type of item (e.g., dichotomous, polytomous), the properties of the item (e.g., discrimination), the number of score levels for the item (in the case of polytomous items), and the distribution of target trait levels in the sample. Thus, there is no single rule to follow for deciding on a requisite sample size to conduct an IRT analysis. Nonetheless, I include this section to provide a general idea of the

required sample size if one intends to use the IRT framework. As a general rule of thumb, the two-parameter model for dichotomously scored items and the GRM for polytomously scored items require at least 500 respondents to have adequately stable estimates of item difficulty and discrimination, although larger sample sizes are desirable and often necessary. Note that this level of sample size is substantially larger than the recommended requirement for classical test theory measures of item difficulty and discrimination. A comprehensive account of the literature investigating IRT sample size requirements is provided by de Ayala (2009).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Blais, A.-R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making, 1*, 33–47.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(Suppl. 9), 635–694.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

CHAPTER 8

BIAS IN PSYCHOLOGICAL ASSESSMENT AND OTHER MEASURES

Jeanne A. Teresi and Richard N. Jones

The focus of this chapter is on methods used to detect item-level measurement bias. Issues surrounding the identification of scale (test) bias are considered briefly. The primary reason for this emphasis is that, as argued later, one cannot legitimately examine scale-level bias without first examining item-level measurement equivalence. Moreover, the methods used for examining scale-level bias have been fairly straightforward and have often involved simple observed variable regression models. Several terms that have been used interchangeably but that have different meanings are *measurement equivalence, bias, invariance, differential item functioning* (DIF), and *fairness*. Formal definitions are given later in this chapter; however, some general terminology is introduced here. *Equivalence* is a broad term that includes conceptual as well as statistical equivalence of items and scales. *Invariance* is a broad statistical term that refers to permissible statistical operations as well as to a hierarchy of statistical models aimed at evaluating the equivalence of item parameters across sociodemographic groups (Meredith, 1993; Rupp & Zumbo, 2004; Teresi, 2006). In general, invariance involves examining conditional relationships. For example, item-level measurement invariance examines the relationship between response to an item and group membership after conditioning (controlling, adjusting) for group differences in level on the trait. Test fairness and bias are viewed as distinct, and *fairness* has been defined as having a social connotation (Society for Industrial and Organizational Psychology, 2003). In general, *psychometric* (*measurement*) *bias* refers to a "systematic inaccuracy of assessment" (Millsap & Everson, 1993, p. 297), such that scores may not mean the same thing in comparison groups.

As reviewed by Meredith and Teresi (2006), the concepts of test fairness and item bias and later DIF arose in the context of educational testing and admission practices. Lord's (1980) contribution brought the topic of testing statistically for item bias to the attention of methodologists, and several seminal reviews followed (e.g., Millsap & Everson, 1993; Potenza & Dorans, 1995). Item bias detection methods arose in part because of the use of high-stakes assessment in education and many branches of psychology and the subsequent concern about the performance of measures across groups differing in sociodemographic characteristics. Because the term *bias* has multiple meanings, some possibly pejorative, and because decisions about items in tests involved expert judgment as well as statistical findings, *differential item functioning* was introduced to refer to the more narrow set of statistical procedures

This work was supported by the following grants: Patient Reported Outcomes Measurement Information System, National Cancer Institute Grant U01-AR057971 and National Institute of Arthritis and Musculoskeletal and Skin Diseases Grant U01-AR057971; Centers of Excellence in Health Disparities, National Institute of Minority Health Disparities Grant P60-MD000206; Resource Centers for Minority Aging Research, National Institute on Aging Grant P30-AG15272-12S2; and the Claude Pepper Older Americans Independence Center, National Institute on Aging Grant P30-AG028741.

This chapter is dedicated to the memory of William Meredith, who provided seminal contributions to the theoretical development of the field of measurement invariance.

DOI: 10.1037/14047-008
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

used to determine whether or not items were performing in the same way across groups (Holland & Wainer, 1993). *Bias* was reserved for a more general finding of group differences, based on expert review and often hypothesis driven, as well as the accompanying body of statistical results known as DIF. Although many of the methods used in the assessment of item-level bias came from the field of educational psychology; this chapter's focus is not on aptitude or achievement tests specifically. Bias analyses have also been central to examining the performance of measures in the fields of industrial and organizational psychology, health psychology, and neuropsychology. Indeed, these psychological measures may also result in decisions that have an impact on the lives of individuals who are assessed.

The chapter is organized as follows. First, background material is presented about the motivation for examining bias in assessments, followed by definitions of key concepts. Test bias is discussed briefly, followed by a discussion of the synergy between item-level DIF analyses, DIF impact, and prediction bias. An illustration is presented. Latent variable methods for assessing DIF are presented, together with magnitude and impact measures. The chapter concludes with a discussion of the meaning of DIF findings and recommendations for future studies examining measurement bias. Throughout, the reader is assumed to be familiar with latent variable modeling concepts in general and item response theory (IRT) in particular (see Chapter 6, this volume).

REASONS FOR ASSESSMENT OF MEASUREMENT BIAS: THE INFLUENCE OF CULTURE

Measurement bias can occur through cross-cultural differences in the interpretation of the meaning of concepts as well as in items used to measure constructs (Rogler, 1989). Language translation and transfer across cultures can have an impact on the psychometric properties of a measure by affecting the difficulty (severity) of the items in the measure, the amount of information provided, and the range of the trait that the measure is able to differentiate (see also Volume 3, Chapter 26, this handbook). Back-translations of standardized items that reflect idiomatic expressions of a particular racial or ethnic group or that have a culture-based meaning may alter the intent of the original item (Angel, 2006). For example, the item "no ifs, ands, or buts" is intended to measure dysarthria (difficulties in repetition of consonants). Because of the inability to identify a colloquial phrase that is a tongue twister, the item is easier in Spanish when literally translated (Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006). As another example, the item "down and blue" has no literal translation in some languages. Finally, Ramirez, Teresi, Holmes, Gurland, and Lantigua (2006) cited an example of very different translations and responses to cognitive items across different groups and studies. They found that Puerto Ricans and Dominicans of different educational backgrounds and ages answered the item "What date of the month is today?" differently. Second-generation, younger Latinos with higher education gave the intended answer (stating the full date). The others responded with the day of the week, interpreting the item in a different fashion.

In addition to different interpretations of translated items, different response styles may result in bias. For example, as reviewed in McHorney and Fleishman (2006), Hispanics of some backgrounds are more likely to use an extreme response style (greater endorsement of the extremes rather than the midpoints of rating scales), and older people have been found to give rosy reports (positive response bias). Such response bias may affect psychological assessments and could have consequences for some individuals. For example, limited proficiency in the majority language of a country has been found to result in substandard health care and deleterious outcomes (Pérez-Stable, 2007); this disparity may in part result from biased assessment resulting from language barriers. Selection and mental health care decisions are often made on the basis of assessments of the cognitive and mental health status of individuals, yet evidence regarding the equivalence of such measures is sparse. To determine whether differences in rates among racial–ethnic, age, and gender groups reflect actual differences and not item bias, studies of factorial invariance and DIF are needed. As an example, reviews of studies of scales assessing depressive

symptomatology have documented evidence of bias for one or more sociodemographic groups, defined in terms of race, ethnicity, education, sex, age or other (Mui, Burnette, & Chen, 2001; Teresi, Ramirez, Lai, & Silver, 2008).

SOCIODEMOGRAPHIC CHARACTERISTICS AND THE GROUP OR STUDIED VARIABLES

Analysis of bias typically depends on selection of a group variable. Of particular interest in educational testing, employment selection, and psychological assessment are race and ethnicity. Several questions arise. Do groups defined, for example, by race, ethnicity, or language constitute homogeneous meaningful entities or are they proxies for other variables? Should group variables such as race and ethnicity be "deconstructed" using factors such as acculturation, educational background, or literacy, including numeracy and reading level? Does race reflect underlying genetic or cultural homogeneity? *Race* is often defined as a social construct based on phenotypic traits such as skin color or hair features. Because groups defined by race have more genetic diversity among them than between them, race as a construct has been argued to lack a biological basis (Manly, 2006). Some have argued against considering ethnic groups monolithically, as has been done in the United States Census and by many government agencies. Racial and ethnic identification data collected on many forms and by the U.S. Census are used by governmental educational and health agencies to track employment, education, health, and housing disparities; however, many individuals no longer self-identify using traditional census categories (Saulny, 2011).

Nonetheless, mental health and other health care providers have viewed the collection of information about race, ethnicity, and language as important with respect to selection of culturally sensitive treatments; however, many are reluctant to ask the questions (Baker et al., 2007). In a report of the National Research Council, Blank, Dabady, and Citro (2004) concluded that although race is a complex social construct, the definition of which is evolving, data on race and ethnicity should continue to be collected and included in policy research. The implication for bias analyses is that race and ethnicity should still be examined; when samples sizes permit, other sources of bias such as education and acculturation should be examined as well. However, Manly (2006) recommended measuring educational quality by asking questions about the region of the country and the level of school segregation. Variables that might also be considered in lieu of ethnicity are place of birth, length of time in a country, age at immigration, main language spoken at home, when the majority language was learned, reading level, literacy, socioeconomic status, and racial socialization. Future work may use different racial and ethnic group designations or attempt to deconstruct these categories. If race is considered a social construct, then it makes more sense to define groups on the basis of the social characteristics described earlier. However, analysis of the numerous possible interactions (among variables; e.g., literacy, acculturation, and education) remains a challenge.

QUALITATIVE METHODS

As reviewed earlier, because translations of instruments can be affected by lack of conceptual equivalence (meaning of terms and constructs) across groups, qualitative analyses are important. Qualitative methods of examining item bias are presented in Volume 3, Chapter 26 of this handbook. These methods, including focus groups, cognitive interviews, and standardized translation processes, are critical to reducing bias in assessments and should be performed before analyses; however, qualitative methods have also been used subsequent to statistical findings of DIF to examine reasons for DIF that may indicate bias, such as changes in format, difficulty of words or phrases, or changes in content that can affect cultural relevance (Angel, 2006; Johnson, 2006; Krause, 2006; Nápoles-Springer, Santoyo, O'Brien, & Stewart, 2006). A best-practice approach to DIF analyses is the generation of hypotheses on the basis of expert review (Roussos & Stout, 1996) and prior findings from the literature of DIF in similar items (Hambleton, 2006). A table can be constructed summarizing hypotheses related to DIF as well as findings of DIF in the literature; an example is provided in the DIF testing of a depression item bank (Teresi et al., 2009). DIF evaluation must be performed in a careful manner that includes

sensitivity analyses using different statistical approaches and examination of the magnitude of DIF (see the section Magnitude and Impact of Differential Item Functioning later in this chapter). Content experts should always be involved in decision making, and item bias analyses should not be viewed mechanistically, as a by-product of statistical procedures.

QUANTITATIVE METHODS: DEFINITIONS AND KEY CONCEPTS

In this section, key concepts related to measurement bias are defined. The section begins with a definition of terms, tracing the origin of the term *measurement invariance* and its distinction from factorial invariance in the factor analysis literature. Statistical evidence for measurement bias alone is not sufficient for an inference that the bias is related to the construct of interest. As discussed in the previous section, qualitative approaches help to clarify why and how measurement differences that are detected might reflect bias in measurement. Additionally, the magnitude and impact of DIF should be evaluated. Although there is limited research on the impact of DIF, simulations have shown that DIF of moderate levels (as contrasted with lower magnitude DIF and smaller numbers of items with DIF) can result in inflated effect sizes associated with observed scale scores (Li & Zumbo, 2009). Such inflation would result in erroneous conclusions about how comparison groups differed on the measure. Many empirical studies have also found that DIF of moderate magnitude can impact scale scores (see Teresi, Ramirez, Jones, Choi, & Crane, 2012).

Measurement Bias

Invariance. *Invariance* is a term that has often been used synonymously with *bias* or *differential item functioning*. This terminology arose out of the factor analysis literature, particularly Meredith's (1964) seminal work. However, measurement invariance, as conceptualized by Meredith (1964), is not the same as factorial invariance, although it is related. Measurement invariance implies that the conditional distribution of a manifest variable, given a value of the latent variable to be assessed, is the same across the groups studied. Because this definition involves conditional distributions of latent variables and is not testable, weaker forms of measurement invariance were introduced. Factor-analytic methods are often used to examine measurement invariance. Meredith (1993) distinguished between strong and strict factorial invariance: Strong factorial invariance is achieved if the conditional expectation of the item response given the common and specific factors is invariant across groups. This level of invariance is tested by examining whether factor loadings and intercepts are invariant. Strict factorial invariance adds the requirement of equivalent group residual variances. When strong factorial invariance is met, group mean differences are comparable across groups; when it is not, one cannot sensibly compare groups on the same latent variable. Meredith argued further that strict factorial invariance is required to ensure fair comparisons between groups; this position has recently been reaffirmed (see A. D. Wu, Li, & Zumbo, 2007, for a review).

Item bias and differential item functioning. The term *differential item functioning* was introduced by Holland and Thayer (1988), disseminated to a wider audience in the seminal volume on DIF by Holland and Wainer (1993) and, as discussed earlier, a distinction was made between bias and DIF. As pointed out by Angoff (1993) in the first chapter in that volume, bias has both a statistical referent (deviance of an estimate from a true value) and a social definition related to the "fairness" of a measure. Thus, DIF was embraced as referring to the statistical findings and bias as referring to a wider process involving hypothesis, expert evaluations, and consideration of collective, cumulative evidence from the literature. DIF involves the evaluation of conditional relationships between item response and group membership. An item shows DIF if people from different subgroups but at the same trait level have unequal probabilities of responding affirmatively to a particular item. A common definition of DIF evolving from IRT is that the item characteristic curves (ICCs) for members of two or more groups are not equivalent. An example is presented later in this chapter. Also see Chapters 6 and 7 in this volume for discussions of IRT and ICCs.

Types of differential item functioning. Two basic types of DIF, described next with respect to each

method, are uniform and nonuniform. As an illustration, a randomly selected woman with low levels of a trait (e.g., perceived psychological distress) should have the same chance of responding in the low-distress direction to an item measuring distress as would a randomly selected man with low distress. For this example, uniform DIF indicates that the DIF is in the same direction for both comparison groups across the distress continuum, whereas nonuniform DIF means that the probability of response in the low-distress direction is higher for men at certain points along the distress continuum and higher for women at other points, thus changing directions.

Magnitude and Impact of Differential Item Functioning

Many items with DIF may be observed, particularly with a very large sample size. It is thus important to consider the practical meaning of DIF. DIF that is of little consequence may be interesting academically, but not clinically or in a utilitarian sense (see Stark, Chernyshenko, & Drasgow, 2004). The core indicators of the practical implications of DIF are magnitude and impact.

Magnitude. *Magnitude* refers to the degree of difference in item performance between or among groups, conditional on the trait or state being examined, and it relates to item-level effect sizes. *Magnitude* has also been defined as the weighted (by the trait distribution) group differences in the probability of an affirmative item response (Wainer, 1993). Magnitude can be measured by examining parameters or statistics associated with the method, for example, the odds ratio. Magnitude measures are important because trivial, nonsalient DIF may result from reliance on significance tests alone.

Impact. Internal impact goes beyond the item level to determine the impact of DIF on the entire measure or scale. Impact can be assessed at the aggregate level by examining group differences in the relationship between the expected scale score and the psychological distress estimate (test response function) or by examining how much mean group differences in total score distributions change with and without inclusion of the items with DIF. DIF may also influence the relationship between self-reported psychological variables and predicted outcomes such as access to care or programs. This latter relationship has been referred to as *external impact, predictive validity,* or *predictive scale bias* and may be examined in terms of predictive values and regression coefficients (see Aguinis, Culpepper, & Pierce, 2010). The impact measures just described are all at the aggregate or group level rather than the individual level. The impact on specific individuals rather than on the group as a whole can also be examined. When selection and treatment decisions are based on individual person assessments, the presence of DIF in the measure can result in bias and negative impact (see Prediction Bias section).

Fairness and Equity

As reviewed in its policy statement on validation, the Society for Industrial and Organizational Psychology (2003) views fairness as a social concept that encompasses factors such as access to information. This concept can be broadened to include access to treatments or resources. More technical definitions are given by Meredith (1993) and Meredith and Teresi (2006), who argued that modeling latent variables is the only way to examine fairness analytically. Because a strict psychometric definition of *fairness* requires unreasonable assumptions about distributional forms, Meredith derived a model of weak fairness, which he defined to exist if the conditional mean vector and variance–covariance matrix of the predictor (x), given any true value of the outcome (y), are identical across comparison groups. In this view, fairness is concerned with decisions to award a resource to individuals before the outcome is known (prospective). Equity involves employment decisions and examination of outcomes after the fact (retrospective; Millsap & Meredith, 1992, 1994).

PREDICTION BIAS

Psychological and educational testing has a long history of examining test bias through regression analysis, specifically predicting performance from a model with inclusion of the test and a group indicator. Test or predictive bias occurs when the intercept and/or slope of a regression line relating the predictor to the outcome are different for different groups.

Methods Used to Examine Test or Prediction Bias

The methodology still used in most test bias studies dates to Cleary (1968), in which the regression equation relates the predictor (e.g., a test) to the criterion (e.g., a rating of success or performance). Typically, the regression analysis is structured as a series of nested models, first entering the predictor variable; the second equation adds the group and Group × Predictor interaction term. The proportion of variance explained is compared between the two models and tested for significance. If significantly more variance is explained when the group and interaction terms are added, then test bias is said to exist. To determine whether bias is due to the intercept or slope, a third model is examined, retaining the group term but removing the interaction term. If the ΔR^2 between the full model (with group and interaction terms) and the reduced model excluding the interaction term is significant, this finding indicates bias in the slopes. If the ΔR^2 between the reduced model (without the interaction) and the first model (with only the predictor) is significant, it indicates bias in the intercept because addition of the group term results in additional variance explained (see Aguinis et al., 2010).

In short, the method that is used in the prediction bias literature is to relate observed assessment measures to observed outcomes. The method implies that the test should be homogeneous; however, homogeneity is not a prerequisite. Additionally, the outcome could be multidimensional. The assessment (X) could be scores on a personality assessment measure, an educational test, a mental health scale, or a measure of health-related quality of life such as those widely used in valuing health and preference measurement in health economics. The outcome (Y) could be performance, grades, dropout rates, diagnosis, subjective well-being, morbidity, and mortality. To avoid confusion, the notation presented next is the same as that used in the test bias literature (e.g., Aguinis et al., 2010; Saad & Sackett, 2002).

$$\text{Model 1: } Y = b_0 + b_1X + e,$$

$$\text{Model 2: } Y = b_0 + b_1X + b_2G + e, \text{ and}$$

$$\text{Model 3: } Y = b_0 + b_1X + b_2G + b_3XG + e.$$

In these models, G is the sociodemographic group variable and XG is the Group × Test Score interaction used in the test of differences in the slope; the beta coefficients are the weights associated with each term; b_0 is the intercept, and e is the error term. Note that this standard effect modification model fails to incorporate the notion of errors in variables, namely, X (test score) = T (true value) + E (error).

In the preceding set of equations, X is assumed to be equal to T. However, measurement error could be estimated in a latent variable model, for example, with an explicit set of equations for the measurement component of the model in addition to the structural equation part linking X to Y. The following measurement model could be specified: $X = \tau_x + \Lambda_x\xi + \delta$, where ξ is the unobserved latent independent variable (representing the underlying latent assessment variable or trait), Λ_x represents the regression weights or factor loadings, and τ and δ are the intercept and error terms, respectively.

Most of the variables examined in prediction studies are not true (error-free) criterion variables, but ratings and other outcomes that are measured with error and can be described as underlying latent variables. However, even in the situation in which a performance or criterion measure is available, it too is measured with error, which may affect the relationship between the test and the performance outcome. Prediction bias analyses are frequently performed using an observed score method and often without examining item-level invariance. Failure to consider the measurement error and to establish measurement equivalence can result in inaccurate estimates of prediction bias. A considerable literature has shown the impact of errors in variables on relationships. This literature has developed in several fields, including psychology, education, mathematics, biostatistics, and econometrics, apparently almost independently, as is evidenced by the lack of citations to work in the parallel areas. In statistics, errors-in-variables is a major topic of interest that evolved on the basis of the work of Carroll and his colleagues (e.g., Carroll, 2003; Carroll, Gallo, & Gleser, 1985; Ma, Hart, Janicki, & Carroll,

2011). Similarly, J. Cohen (1988) in psychology and Fleiss (1986) in biostatistics showed the relationship between reliability and statistical power and the effect of unreliability on structural relationships. As is well known, regression estimates are affected by errors in variables; such errors can be induced by several factors, including DIF.

The body of applied research based on the test bias model just described resulted in the conclusion that bias in the slopes is usually not observed, but that intercept bias is often observed. As reviewed by Aguinis et al. (2010), these studies may be flawed by several factors that affect the power to detect differences in the slopes: (a) The sample sizes are too small in most studies; (b) the proportion of subgroup (usually ethnic minority) sample members is too small and the subgroups are unequal; (c) the ranges are restricted; and (d) unreliability both affects the power of the studies to detect slope differences and increases the Type I error, resulting in false prediction bias in the intercepts (see Aguinis et al., 2010). To this set of potential problems with the studies of prediction bias, it could be added that measurement error is often not evaluated, and item-level invariance is often not established.

In the next section, an argument is advanced that unless lower levels of invariance are established, the examination of the predictive invariance is questionable. Scale means cannot be compared legitimately if invariance at the intercept level is not established (e.g., Gregorich, 2006; Meredith, 1993; Vandenberg & Lance, 2000). To account for this error properly, a latent variable model can be used to model the measurement error at the same time (or before) examination of the structural (regression) coefficients representing the relationship between the exogenous and endogenous variables. Also note that the correction for attenuation applied in some structural equation models to adjust for unreliability can also produce inflated estimates of structural coefficients (see P. Cohen, Cohen, Teresi, Marchi, & Velez, 1990), and the quality of measurement (reliability) may also affect model fit in the structural part of the model (Hancock & Mueller, 2011). Thus, the models must be specified carefully. It is also acknowledged that latent variable model trait estimates based on fixed-length scales may not perform in a superior manner to observed scores (see Xu & Stone, 2012). Nonetheless, as reviewed in the next section, there is an important role for latent variable models in studies of prediction bias. In summary, the analyses in many studies are performed using simple regressions in examining predictive bias rather than latent variable models. Therefore, the capacity to adequately model the relationship between observed variables that are measured with error is limited (see B. O. Muthén & Hsu, 1993).

Differential Item Functioning Testing Versus Differential Item Functioning Impact, Prediction Bias, and Validity

At a National Institutes of Health (NIH) Patient Reported Outcomes Measurement Information System (http://nihpromis.org) investigator meeting that included National Institutes of Health scientific committee project officers with backgrounds in industrial psychology, an opinion was expressed that DIF analyses were not required because it had been shown that all that was necessary was to establish predictive validity. This argument follows from Hunter and Schmidt's (2000) often quoted article in which they asserted that bias in professionally developed tests is unlikely because equivalent performance outcomes are observed, given equivalent test performance (see also Sackett, Schmit, Ellington, & Kablin, 2001):

> The study of potential racial and gender bias in individual test items is a major research area today. The fact that research has established that total scores on ability and achievement tests are predicatively unbiased raises the question of whether there is in fact any at the item level. (Hunter & Schmidt, 2000, p. 151)

However, a body of research, summarized by Millsap (2007a), has shown that even if predictive invariance exists, in the presence of a lack of measurement invariance (DIF) systematic selection errors can result when the measure is used in clinical decision making. As reviewed by Millsap, in the view of many influential researchers in organizational and industrial psychology, the question of bias in tests has already been answered, and it does

not exist. For example, investigators such as Sackett et al. (2001) reported that predictive invariance for preemployment tests at the scale level has been well established for Blacks and Whites. However, score differences between Whites and Blacks on cognitive tests remain, and this difference is not explained by factors such as stereotypic threat (Steele & Aronson, 1995) alone, as has been widely reported in the press (Sackett, Hardison, & Cullen, 2004). Stereotypic threat was measured in a series of experiments in which majority and minority high-performing students were presented with an intelligence or achievement test under conditions of threat (e.g., indicating their race or being told the nature of the test) or nonthreat (asked to help in the evaluation of a new problem-solving test). Blacks who believed that the test was a research tool performed significantly better than Blacks in the threat condition (Steele & Aronson, 2004). However, the White–Black gap in test scores remained, leaving unanswered the question of what educational quality and economic factors may contribute to this gap (Sackett et al., 2004; see also Chapter 36, this volume). If the possibility of bias in professionally developed tests as a potential explanation for performance disparity is summarily dismissed, then efforts to understand and remediate such differences between gender and racial and ethnic groups could be hampered.

In another widely cited article, Borsboom (2006a) also referred to this thinking, lamenting the state of affairs in psychology, in which theoretical discoveries of psychometricians (e.g., Millsap, 1997) related to development and evaluation of measures are not integrated into psychology and, the authors would argue, into many other fields that rely on measurement of inherently latent variables. In that article, two important points were reiterated, namely that classical test theory–derived methods and statistics are still the most widely used and reported and that the assumptions common to all observed score regressions used in studying test bias (prediction models) are based on the notion that the variables are measured without error. This practice continues to occur in part because, as described by Borsboom, treating variables as observed and without measurement error permits smaller sample sizes. Clark (2006) noted in one of the accompanying commentaries that the theoretical work showing the need to examine item-level bias is in the context of a specific (single common factor) model; however, such models are appropriate for many predictive bias studies. The key point remains that prediction invariance does not imply measurement invariance or lack of test bias (Borsboom, 2006b).

As just reviewed, the effects of unreliability on linear prediction of observed scores is well known. A parallel literature has examined the effects of unreliability on prediction of criterion variables. In the case of diagnostic observed variables, the relationship between the screening test and the diagnostic outcome is often studied, and the question as to whether the results can be generalized to groups differing in gender, age, race, and ethnicity is asked. The theory behind comparing the performance of a test against a criterion variable (outcome) is that conditional on a test, outcomes are not dependent on the group designation. For example, in the literature on biomarkers, this is known as the *Prentice criterion* (Prentice, 1989) and implies that the test (biomarker) contains all the information about the treatment (or age or ethnic group) effect on the outcome. Thus, the positive and negative predictive values do not depend on group assignment; however, this criterion may not hold in practice (Dodd & Korn, 2008). One can only generalize summary statistics such as predictive values and sensitivity and specificity to samples with similar distributions and characteristics. Moreover, in the case of observed variables, a large literature in epidemiology and biostatistics has demonstrated the effects of base rate (prevalence) and unreliability on a test's sensitivity and specificity, reinforcing the need to examine a test's properties before moving to predictive validity.

Borsboom, Romeijn, and Wicherts (2008) extended this work to latent variable models in a series of simulations showing the relationship of measurement and selection invariance. As has been shown for observed variables, selection invariance (studied in terms of sensitivity, specificity, and prediction values) is affected by unreliability and the selection ratio; additionally, group differences in the mean latent trait also affect selection outcomes. Borsboom et al. made the important point that measurement and prediction invariance are not the same thing and that invariant

test criterion regressions should not be viewed as evidence of measurement invariance.

MEASUREMENT BIAS: METHODS FOR EXAMINATION OF DIFFERENTIAL ITEM FUNCTIONING

As discussed at the beginning of the chapter, all methods for DIF detection include a conditioning variable because it is important to compare the comparable (see Dorans & Kulick, 2006). Group differences in response to an item may occur because of actual differences in the distribution of the trait in the groups. For example, men and women may differ in overall levels of depression, resulting in mean differences in the estimate of the depression trait and different endorsement rates for items. These differences in rates do not automatically translate to DIF. The items require testing for group differences after conditioning on the trait. The question then is "Does the likelihood of item endorsement differ between groups, given that individuals are at the same level of the trait measured"?

Because studies have shown that the conditioning variable should be free of DIF, most DIF detection measures require a set of DIF-free items, often called an *anchor set,* to link the groups on the trait. The assumption is that the anchor items perform similarly in both groups and that they are potentially the set that forms a scale with a common metric for the two comparison groups (see Orlando Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006). *Purification* is a procedure in which anchor items that are DIF free are identified through an iterative process. For example, Mazor, Hambleton, and Clauser (1998) proposed a two-stage DIF evaluation: First, all test items are examined for DIF; next, those items showing DIF are removed, and the process is repeated. The best method of selecting these anchor items has been the subject of numerous reviews (e.g., Wang & Shih, 2010; Woods, 2009a) and is beyond the scope of this discussion.

Nonparametric Methods

Several nonparametric methods have been used to examine DIF: Mantel–Haenszel (1959; Holland & Thayer, 1988), standardization (Dorans & Kulick, 1986; Dorans & Schmitt, 1993), and simultaneous item bias test (Chang, Mazzeo, & Roussos, 1996; Stout & Roussos, 1995). These methods are not discussed here because numerous reviews of such methods exist (e.g., Millsap & Everson, 1993; Potenza & Dorans, 1995) and because, as discussed earlier, latent variable models provide a more flexible framework for DIF detection. Finally, simulation studies, for example, Woods (2011), are beginning to provide evidence of the equivalent or superior performance of latent variable parametric models compared with nonparametric models under several conditions that have been found to adversely affect DIF detection methods.

Parametric Methods

The parametric observed score methods include logistic and ordinal logistic regression (Swaminathan & Rogers, 1990; Zumbo, 1999). Parametric latent variable methods include IRT ordinal logistic regression (IRTOLR; Crane, van Belle, & Larson, 2004), a logistic regression approach with a latent conditioning variable; multiple indicator–multiple cause (MIMIC; B. O. Muthén, 1984; L. K. Muthén & Muthén, 2012); analyses of residuals from the Rasch model (Rasch, 1960; Wright & Stone, 1979); IRT log-likelihood ratio (Thissen, 1991, 2001; Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1993), and differential functioning of items and tests (DFIT; Raju, 1999; Raju, van der Linden, & Fleer, 1995). Also briefly introduced are some new approaches. Because the models have been presented elsewhere, only brief descriptions are provided. Discussion of several factors that influence DIF detection, including purification, model fit and assumptions, trait distributions, sample size, and cutoff values associated with magnitude measures, is informed by the results of simulation studies. It is beyond the scope of this chapter to present the findings; however, reviews, including advantages and disadvantages, can be found in Teresi, Stewart, Morales, and Stahl (2006).

Examples of Differential Item Functioning

Figure 8.1 displays a few examples of different types of DIF. Plots represent the ICC for a reference group (solid line) and a focal group (dashed line) as well

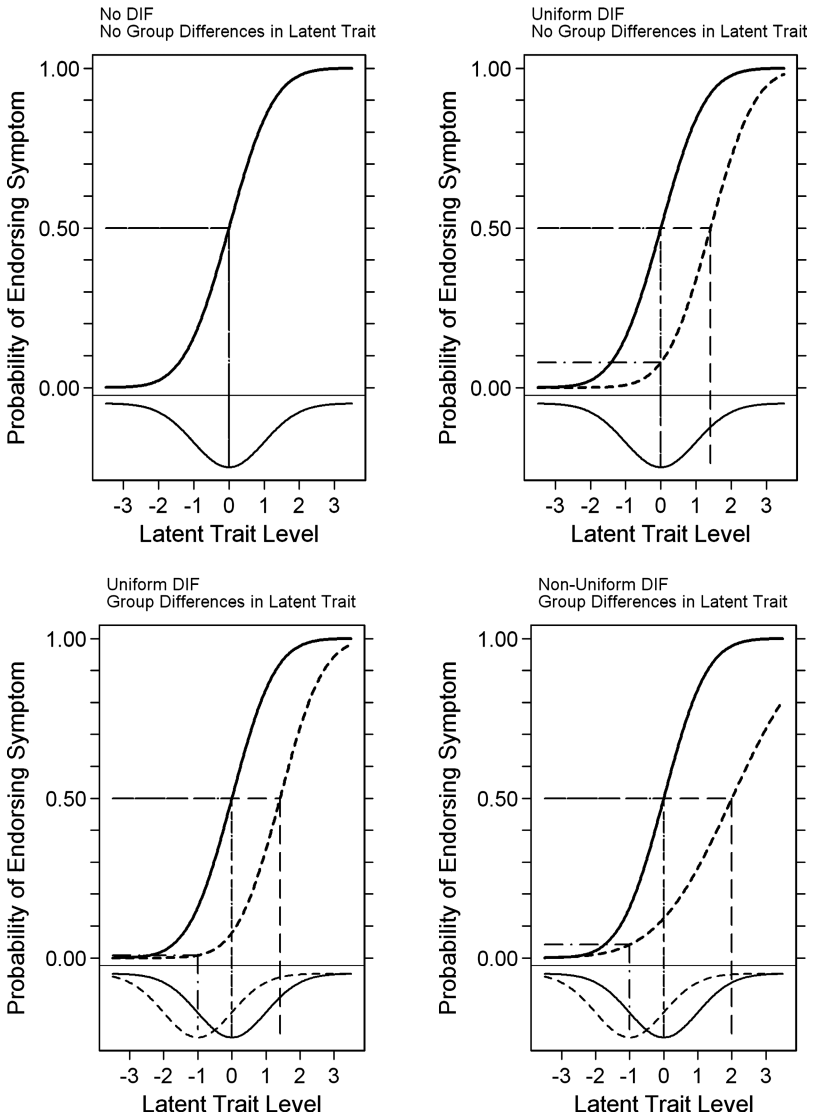

FIGURE 8.1. Types of differential item functioning (DIF). Solid line = reference group; dashed line = focal group; vertical dashed line = symptom severity; horizontal dashed line = expected proportion of individuals in the population who will endorse the symptom.

as the complement density illustrating the distribution of the latent trait in the two groups. (The focal or studied group is usually the group that is the target of investigation, as described in the beginning of the chapter.) Although the plots are of binary items, similar plots could be constructed for polytomous items. In the top left panel, there is no DIF and there are no group differences in the distribution of the latent trait: The two groups' ICCs lie atop one another, as do the curves representing the latent trait density. The vertical dashed line illustrates the symptom severity: the level of the latent trait at which a randomly selected person from the population has a 50% chance of endorsing the symptom. The horizontal dashed line indicates the expected proportion of individuals in the population who will endorse the symptom, which is also 50% because in this example the symptom severity is perfectly matched to the trait level in the group.

The top right panel of Figure 8.1 illustrates uniform DIF: The item has a higher threshold in the focal group. The severity for this item is about 1.4 in the focal group and is still 0 in the reference group. In this example, the two groups continue to have

equivalent latent trait distributions. However, because the symptom has a greater severity for the focal group, members of this group must have a higher level of the latent trait before a randomly selected person has a 50% probability of endorsing the symptom. The probability of endorsing the item is greater for the reference group than for the focal group at all levels of the latent trait.

In the bottom left panel, group differences in the latent trait are added to the illustration of uniform DIF. The symptom has a higher severity threshold in the focal group, but the focal group also has a much lower average level on the latent trait. The expected proportion of focal group members who endorse this symptom is nearly 0, compared with 0.50 in the reference group. Finally, the bottom right panel illustrates nonuniform DIF. It shows a situation with group differences in the distribution of the latent trait, differences in item severity threshold, and group differences in the strength of the association of the item and the latent trait, or symptom discrimination, which is proportional to the slope (see Chapter 6, this volume). It should be noted that typically nonuniform DIF is illustrated by curves that cross; however, they may cross beyond the range of the trait depicted in the diagram.

As an example, if one found that women are more likely to endorse symptoms of tearfulness than are men, this type of DIF might be represented by the curves in the lower left or lower right panel, under the assumption of true sex differences in the distribution of the latent depressive trait. DIF detection techniques are available to identify and quantify possible bias introduced by divergent item parameters in two groups as distinct from differences between groups on the latent trait.

Evaluating Model Assumptions: Dimensionality and Local Independence

Many of the methods that have been used to examine DIF assume local independence and essential unidimensionality. Local independence implies that items are not correlated, after conditioning on the trait. Tests of local independence, for example, Chen and Thissen (1997) and Orlando and Thissen (2003), are incorporated into various software packages described in Appendix 8.1. Such tests and IRT model fit are not discussed here, but issues related to model fit in the context of structural equation models is presented briefly in the Item Response Theory, Factor Analyses, Differential Item Functioning, and Invariance section (see also Millsap, 2007b). Unidimensionality implies that one latent trait underlies the data. Violations of the unidimensionality assumption can adversely affect DIF detection results (e.g., Ackerman, 1992; Mazor et al., 1998; Snow & Oshima, 2009). Approaches to assessing dimensionality include nonparametric methods (e.g., Stout, 1987) and parametric methods, for example, analysis of standardized residuals (Y.-T. Chou & Wang, 2010). Methods for modeling dimensionality and estimating reliability have advanced, and it is important to apply these advances. For example, examining dimensionality with exploratory factor analyses of polychoric correlations is described in several articles (Reise, Cook, & Moore, in press; Reise, Moore, & Haviland, 2010; Reise, Moore, & Maydeu-Olivares, 2011; Reise, Morizot, & Hays, 2007) and in Chapter 6 in this volume. To assess the degree to which the scale measures one common factor, Revelle and Zinbarg (2009) recommended the use of factor analyses with a Schmid-Leiman (1957) transformation for exploratory factor analyses, with subsequent estimation of omega hierarchical (ω_h, a reliability estimate) based on confirmatory factor analysis (CFA), and use of a bifactor model. Although the exploratory bifactor analysis can be conducted with Schmid-Leiman orthogonalization, it is recommended that final loadings be estimated with other modeling procedures (Reise et al., 2011; for alternative exploratory bifactor models, see See Jennrich & Bentler, 2011, 2012). The explained common variance provides information about whether the observed variance–covariance matrix is close to unidimensionality (Sijtsma, 2009). It can be estimated as the percentage of observed variance, calculated as the ratio of the first eigenvalue to the sum of all eigenvalues extracted from a bifactor model analysis (see Reise et al., 2010).

Description of Differential Item Functioning Detection Methods Based on Observed Variable Models

Logistic regression DIF methods (Swaminathan & Rogers, 1990) provide tests of whether the conditional

odds of endorsing a symptom are different between two groups. For each item, a set of regression equations is developed, predicting item scores from raw trait scores, group membership, and the Group × Trait Score interaction.

A general model is $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3(x_1 x_2)$, where y is the item response variable, x_1 is the trait variable, β_1 is the coefficient, x_2 is the group or studied covariate, and β_2 is the coefficient, which for some models represents the difference in item performance between groups that can be exponentiated to obtain an odds ratio. This group term is an estimate of uniform DIF. If this coefficient is not significant, the result supports the hypothesis of no DIF because the trait scores explain all of the variance in items scores. β_3 is the coefficient for the Group × Trait interaction. The interaction term $\beta_3(x_1 x_2)$ is an estimate of nonuniform DIF. Polytomous response data can be modeled using ordinal logistic regression (Zumbo, 1999), in which the item response y is specified as a latent continuously distributed random variable. The magnitude of the overall DIF can be computed by subtracting the R^2 value for models with only the trait term from that for models with all terms. Gelin and Zumbo (2003) used Jodoin and Gierl's (2001) two criteria that must be met for an item to be classified as displaying DIF. First, the two degrees-of-freedom chi-square test for DIF (testing for the group and interaction effects simultaneously) must have $p \leq .01$. Second, the corresponding measure of effect size (R^2) must have a value of at least .035.

Crane et al. (2004; Crane, Gibbons, Jolley, & van Belle, 2006) proposed the substitution of IRT-based latent trait variables (θ) for observed conditioning variables. Nonuniform DIF is determined via a difference in −2 log-likelihood (distributed as a chi-square) for nested models with and without inclusion of an Ability × Group interaction term. Uniform DIF is identified on the basis of the change in the beta coefficient for ability when a group term is added to the model. An estimate is calculated of the difference between the beta for the reduced or compact model and that of the augmented model divided by the beta for the compact model. If the value for the ratio is greater than an empirically derived cutoff value based on the data set examined, one may conclude that uniform DIF is present after ruling out nonuniform DIF (Choi, Gibbons, & Crane, 2011).

Item Response Theory Models and Differential Item Functioning Tests Derived From Item Response Theory Latent Variable Models

IRT approaches are one application of general latent variable modeling. IRT can be used to introduce a latent variable as the conditioning variable, which adds considerable strength in terms of drawing inferences about possible measurement bias. However, these gains do come at the cost of increased computational complexity. The sections that follow describe popular IRT latent variable models that are used for DIF detection. They include the two-parameter logistic response model for binary data and the graded response model for polytomous items. Three DIF detection methods and statistical approaches based on these models are also described.

Rasch model. The basic Rasch model approach used in many studies of DIF is a latent variable–based method that compares the item locations (difficulties, severity) between two groups (Wright & Stone, 1979). It does not incorporate modeling of the discrimination (slope) parameter. This model was originally restricted to binary data; however, an extended Rasch model that incorporates polytomous data has been used in DIF testing in the context of fit (e.g., Hagquist & Andrich, 2004). For reviews, see Mair and Hatzinger (2007b) and Kubinger (2005). An important consideration in the use of Rasch-based methods is whether the assumption of equal discrimination parameters holds for psychological and mental health data.

Graded response model. The graded response model (Samejima, 1969) is used in many IRT log-likelihood tests of DIF in polytomous items (e.g., Thissen, 1991; Thissen et al., 1993). Ordered responses, $x = k$ and $k = 1, 2, \ldots m$ are assumed, in which a_i is the discrimination (slope) and b_{ik} is the difficulty parameter for response category k: $P(x = k) = P^*(k) - P^*(k + 1) = 1 / \{1 + \exp[-a_i(\theta - b_{ik})]\} - 1 / \{1 + \exp[-a_i(\theta - b_{ik+1})]\}$.

$P^*(k)$ is the ICC describing the probability that a response is in category k or higher for each

value of θ (see also Orlando Edelen et al., 2006; Thissen, 1991). Various magnitude and impact measures are derived from this basic IRT model, as are the DFIT indices described in the section Differential Functioning of Items and Tests later in this chapter.

IRT log-likelihood ratio methods. The likelihood ratio test compares the likelihood of nested models (Thissen et al., 1993). A first step is to construct a DIF-free anchor set of items by testing items one at a time for DIF. As reviewed in Orlando Edelen et al. (2006) and Teresi et al. (2009), a compact (or more parsimonious) model is tested with all parameters constrained to be equal across groups for a studied item (together with the anchor items that are DIF free) against an augmented model with one or more parameters of the studied item freed to be estimated distinctly for the two groups. The procedure involves comparison of differences in log-likelihoods (−2 log-likelihood; distributed approximately as chi-square) associated with nested models; the resulting statistic is evaluated for significance with degrees of freedom equal to the difference in the number of parameter estimates in the two models. Group differences in severity (*b*) parameters are interpreted as uniform DIF only if the tests of the *a* parameters are not significant; in that case, tests of *b* parameters are performed, constraining the *a* parameters to be equal. The final *p* values are often adjusted using methods such as Benjamini-Hochberg (1995; Thissen, Steinberg, & Kuang, 2002).

Wald tests. The Wald statistic is equivalent to Lord's (1980) chi-square, which was extended for polytomous data by A. S. Cohen, Kim, and Baker (1993). The Wald statistic is also asymptotically equivalent to the likelihood ratio test. This method was proposed for use with cognitive assessment data by Teresi, Kleinman, and Ocepek-Welikson (2000); however, more advanced estimation procedures (Cai, 2008) were introduced by Langer (2008) and incorporated into IRT for patient-reported outcomes (Cai, du Toit, & Thissen, 2009; Cai, Thissen, & du Toit, 2012), described in the Appendix. As summarized in Teresi (2000), Lord (1980, p. 223) proposed a chi-square statistic, $\chi^2 = \underset{\sim}{\mathbf{v}}'_i \Sigma_i^{-1} v_i$, simultaneously testing the hypotheses that the *a*s and *b*s of Group 1 on Item *i* are equal to the *a*s and *b*s of Group 2, where $\mathbf{v}'$ is the vector $\left\{ \hat{b}_{i1} - \hat{b}_{i2}, \hat{a}_{i1} - \hat{a}_{i2} \right\}$ and Σ_i^{-1} is the inverse of the asymptotic variance–covariance matrix for $\hat{b}_{i1} - \hat{b}_{i2}$ and $\hat{a}_{i1} - \hat{a}_{i2}$. Because $\hat{a}_{i1}$ and $\hat{b}_{i1}$ are independent of $\hat{a}_{i2}$ and $\hat{b}_{i2}$, $\Sigma_i = \Sigma_{i1} + \Sigma_{i2}$, where Σ_{i1} is the sampling variance–covariance matrix of $\hat{a}_{i1}$ and $\hat{b}_{i1}$, and similarly for Σ_{i2}.

Differential functioning of items and tests. Also based in IRT is DFIT (Flowers, Oshima, & Raju, 1999; Raju, 1999; Raju et al., 1995; Raju, Fortmann-Johnson, Kim, Morris, Nering, & Oshima, 2009). The DFIT methodology permits examination of the magnitude of the gap between the ICCs for two groups, such as illustrated in the lower right panel of Figure 8.1. Noncompensatory DIF is an effect size measure that is weighted by the focal group density such that more weight is given to differences in the region of the trait with the highest frequency in the targeted group. Each respondent is posited to have two true (expected) scores, one as a member of the focal (studied) group and one as a member of the reference group. For item *i*, NCDIF reflects the average (expected value) of the squared difference between expected item scores for individuals as members of the focal group and as members of the reference group (see also Morales et al., 2006; Teresi et al., 2007). For binary items, expected (true) scores equal the probability of item endorsement, conditional on trait (θ) level: $P_i(\theta_s)$. For example, for a binary item, $P_{if}(\theta_s)$ is the expected score for individual *s* as a member of the focal (studied) group, reflecting the probability of a correct response to a test item. Because simulation studies have found overidentification of DIF with the use of chi-square tests, cutoff values are used instead to identify DIF (see Morales et al., 2006). DFIT yields both magnitude and impact measures (see Magnitude and Impact for Item Response Theory–Based Differential Item Functioning Methods section).

Item Response Theory, Factor Analyses, Differential Item Functioning, and Invariance

IRT and confirmatory factor analyses constitute two general methods for the examination of item

invariance. The general relationship between factor analyses and IRT and their equivalence has been the topic of numerous articles (e.g., McDonald, 2011; Mellenbergh, 1994; Reise, Widaman & Pugh, 1993; Takane & de Leeuw, 1987) and is discussed in the next section.

Multiple indicator–multiple cause and mean and covariance structure analysis for differential item functioning detection. Mean and covariance structure models are a special application of structural equation modeling (SEM) that involves multivariate analysis of both covariances and means among multiple dependent variables. A special kind of SEM, MIMIC, is particularly relevant for describing the relationship of a grouping variable, an underlying latent trait, and an item response in a simultaneous system of equations. MIMIC can be a powerful approach for detecting DIF (Woods, 2009b). A more flexible but complex approach is multiple-group CFA, which includes multidimensional models.

Given a set of ordinal item responses, a unidimensional CFA model estimated on a matrix of polychoric correlation coefficients with uncorrelated measurement errors implements a latent trait measurement model that is equivalent to a graded response IRT model (Jöreskog & Moustaki, 2001; Mislevy, 1986). The measurement model can be represented by $y^* = \Lambda\eta + \varepsilon$, where η represents one or more latent variables underlying the item responses. When a model contains one latent variable, the meaning of η is identical to that of θ in other presentations of the IRT model. The outcome variable vector, y^*, is an array of latent response variables underlying the observed and discrete responses, y. The y^* and y variable vectors have a threshold relationship, where y_j is in category c if y_j^* is greater than threshold τ_c and less than or equal to τ_{c+1}. Λ contains a matrix of linear regression parameters, λ, that are factor loadings (analogous to IRT slopes) and describe the per-unit increase in y_j^* per unit increase in η. IRT discrimination parameters (a) can be determined from the factor analysis results in a single factor model using $a_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}}$ under the standard normal latent trait assumption (see also Lord & Novick, 1968). Boundary (difficulty or severity) parameters are $b_j = -\tau_{jc}\lambda_j^{-1}$ (B. Muthén & Asparouhov, 2002).

A single-group CFA model performed within the context of SEM can be extended to allow the inclusion of covariates (x). The measurement model is $y^* = \Lambda\eta + Kx + \varepsilon$, and a second-level system of equations, $\eta = \alpha + \Gamma x + \zeta$, is introduced. In this second level, the so-called structural regression model, Γ, contains regressions of the underlying trait and describes the effects of covariates (background variables, a grouping variable) on the underlying trait. In the first-level model, K contains regressions of the latent response variables on background variables. As such, these effects describe possible (uniform) DIF.

Multiple-group confirmatory factor analyses. The general latent variable modeling approaches (B. O. Muthén, 2002) can be used to extend the CFA model to multiple groups to detect DIF. For example, a measurement model can be estimated separately, but simultaneously, for men and women. Model identification and measurement model calibration are achieved by imposing equality constraints on the measurement model parameters (Λ, τ) and variance parameters for the latent trait across groups. Uniform DIF can be detected by relaxing equality constraints on threshold parameters, and nonuniform DIF can be detected by relaxing equality constraints on factor loadings. Iterative procedures for model building in a forward-stepwise fashion using model modification indices (chi-square scaled derivatives from the model fit function) have been described (Jones, 2006; B. Muthén, 1989a). Covariates can also be entered into multiple-group CFA models. Conventionally, effects in **K** and Γ are assumed to be equal across group, but substantively motivated tests of such assumptions can be performed in a data-driven model-fitting procedure (Jones, 2006). The general analytic approach is a multiple-group structural equation model with mean structures. The SEM approach to detecting DIF has been explicated by B. Muthén and Lehman (1985) and Thissen et al. (1993). Applications can be found in Jones and Gallo (2002); Gallo, Anthony, and Muthén (1994); and Christensen et al. (1999).

The relationship between the MIMIC model and the multiple-group confirmatory factor analysis

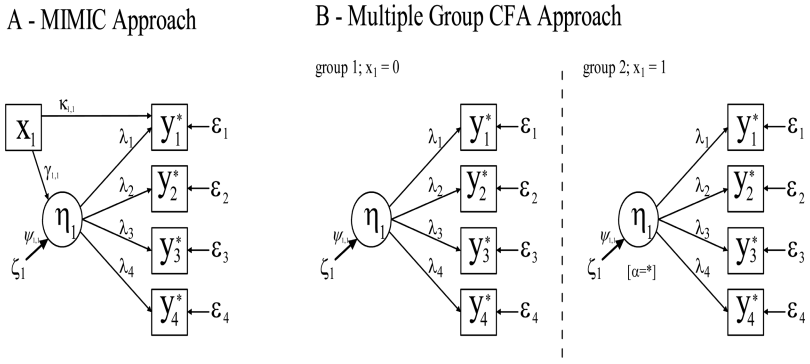

FIGURE 8.2. Multiple indicator–multiple cause (MIMIC; A) and multiple-group confirmatory factor analysis (CFA; B) approaches.

approach is illustrated in Figure 8.2. In the MIMIC model approach (Figure 8.2A), a measurement model relates the latent trait to the latent response variables. The MIMIC model also includes a structural model part, in which the latent variable is regressed on a covariate (*x*). Such regressions are termed *indirect effects* insofar as they capture the relationship of causes to indicators indirectly, via the latent variable. The structural model part also includes a regression of individual test items' latent response variables on covariates (***K***); such effects are termed *direct effects*. In the MIMIC model, a significantly nonzero value for ***K*** is sufficient to describe uniform DIF, an item difficulty shift for members of the group marked by ***x***. Significant indirect effects in Γ capture group heterogeneity.

The single-group MIMIC model as illustrated in Figure 8.2A has limited capacity[1] to capture nonuniform DIF. However, Woods and Grimm (2011) developed a method that examines nonuniform DIF in the context of interaction terms. The single-group model is expanded to include the Group × Trait interaction; however, the interaction term must be carefully constructed, and some assumptions, for example normality, may be violated (see Woods & Grimm).

In the multiple-group CFA model (Figure 8.2B), much more flexibility is permitted. It is possible to relax assumptions of equality on all model parameters, including factor loadings, mean structures, variances, and residual variances. Even the number of latent factors can vary across group. Different levels of equality constraints across these model parts constitute a hierarchy of factorial invariance (e.g., Meredith, 1993). Strong factorial invariance is assumed if groups have equivalent τ (threshold or difficulty) and λ (factor loading) values. In general, group heterogeneity is captured by relaxing equality assumptions on the mean of the latent trait across groups. Uniform DIF is captured by relaxing assumptions of equivalence across group in the means for the latent response variables or thresholds for observed categorical variables. Nonuniform DIF can be captured in relaxing assumptions of equivalence of item factor loadings across group. However, it is important to note that across-group ICCs can be nonparallel if groups differ in any element of the factor loadings, variances for the latent trait, and residual variances for the items. B. Muthén (1989b) has therefore argued that the definition of nonuniform DIF in terms of nonparallel ICCs is overly restrictive. For further discussion of the levels (hierarchy) of factorial invariance, including issues related to residual invariance, see Meredith (1993) and Gregorich (2006).

[1]It is possible to model nonuniform DIF in a single-group MIMIC model using missing data techniques. Essentially, if y_1 is the studied item, the data set is augmented by creating a new variable holding observed values for y_1 for members of the reference group and another copy of the variable holding observed values on y_1 for members of the focal group. Nonobserved values are missing. Thus, a test of four items would have five items when augmented. Uniform and nonuniform DIF is assessed by relaxing equality constraints on factor loadings and mean structures for the new y_1-derived variables. Such models require the use of robust full information maximum likelihood estimator algorithms (e.g., Mplus or MLR) because the low covariance coverage resulting from augmentation will cause limited-information weighted least squares approaches to fail to return parameter estimates. Model selection algorithms for such an approach have not been described, but one could adapt algorithms developed for other platforms.

Magnitude and Impact for Item Response Theory–Based Differential Item Functioning Methods

Because significance tests alone are subject to chance findings, and with large sample sizes trivial differences in item functioning between groups may be significant, effect size or DIF magnitude measures are often used in conjunction with statistical tests in addition to corrections for multiple comparisons (see Teresi, 2006). The impact of item-level DIF on the scale is also evaluated. Examination of magnitude is important when making decisions about whether to remove an item from a measure or to consider providing separate calibrations for different groups. IRT-based indices of DIF magnitude and impact are derived from expected item and expected total (test) response functions. Group differences in the expected proportion correct (for binary items) or the expected item score (for polytomous items) are magnitude measures. The item-level expected score is the sum (over categories) of the probability of response in category *k*, weighted by the category score (e.g., the ordinal code for the category). The total scale expected score or true score can be expressed as the sum over items of the conditional probability of response, and it is a measure of DIF's impact on the entire measure. Measures based on this method are described in S.-H. Kim, Cohen, Alagoz, and Kim (2007) and Teresi et al. (2007) and incorporated into software such as lordif (described in Appendix 8.1 and in Choi et al., 2011). Expected item and scale scores are central in the area statistics and DFIT methodology developed by Raju and colleagues (Raju, 1988, 1999; Raju et al., 1995, 2009). For binary items, the exact-area methods compare the areas between the item response functions estimated in two different groups; A. S. Cohen et al. (1993) extended these area statistics for the graded response model. The averaged unsigned area difference between the expected item response functions weighted by the studied group density, evaluated at various quadrature (θ) points, provides an effect size measure (see Raju, 1988; Wainer, 1993; Woods, 2011). This value is similar to the noncompensatory DIF index (Raju et al., 1995) described earlier, which provides a magnitude measure with respect to item-level data. Differential functioning at the test level is the sum of differential functioning at the item level and indicates how much each item contributes to differential functioning at the test level of the whole. DIF in one item can cancel out DIF in another item.

MIMIC models examine the magnitude of DIF through examination of the direct effect and impact by comparing the estimated group effects in models with, and without, adjustment for DIF (Jones, 2006; Jones & Gallo, 2002). Changes (before and after DIF adjustment) in group differences in mean scores on the latent variable can be used to examine impact.

In addition to aggregate-level impact, individual impact can be assessed by several methods. In the context of latent variable models, individual impact can be examined by fixing and freeing parameters based on DIF findings and examining changes in trait scores. In summary, as stated earlier, examination of magnitude and impact is central to DIF analyses.

Differential Item Functioning Detection With Longitudinal Data

Sources of invariance may be temporal differences in the meaning or interpretation of the construct or response scale over time (Golembiewski, Billingsley, & Yeager, 1976). For example, in summarizing the 40-year follow-up of the participants in the Stirling County Study, Murphy, Laird, Monson, Sobol, and Leighton (2000) described a shift in the thinking about depression symptoms and their relevance over time. They discuss that from 1952 to the 1990s, the prevalence of reporting low spirits declined, and other ways of reporting depressed mood, for example as feeling low and helpless, increased. Such differences in vernacular meaning of items confound attempts to address research questions regarding age differences and cohort differences. Applied researchers, when faced with such a disparity of measurement over time, might be led to discard such items in their longitudinal assessment of depressive symptoms. The use of IRT models and assumptions of partial measurement invariance (i.e., allowing a small set of items to have different measurement properties over time) would permit the retention of items across waves. Differences in the measurement

model are built into the analytic model examining change over time in level of depression, and dependency of measures over time can be modeled, for example, by including a random effect term to model the repeated measures in Cai's (2010) two-tier method (see also Maydeu-Olivares & Coffman, 2006). Further discussion of methods for longitudinal invariance is beyond the scope of this chapter but is presented in the applied literature on SEM by McArdle, Fisher, and Kadlec (2007; see also Chapter 12, this volume).

DISCUSSION

The absence of measurement bias is a prerequisite for valid inferences regarding group differences and individual change over time. However, the great many approaches to DIF detection and the absence of clear guidance in the field regarding magnitude and impact are limiting factors in the integration of measurement bias research in applied research settings.

Meaning and Importance of Differential Item Functioning

Despite several well-characterized statistical procedures for detecting DIF, the field is without practical guidance on what magnitude of DIF is important, and what to do about it if DIF is determined to be caused by item bias. A fundamental problem with DIF statistics is that they represent group-level effects. Whether the presumed mechanism that causes DIF to appear in aggregate analyses of group differences is at work in a particular person drawn from the group is unknown. A widely disregarded aspect of DIF analyses concerns the level of precision of the instrument used to measure the trait. Practical and resource considerations in applied clinical and research settings put downward pressure on the length of measurement instruments. In general, the shorter the instrument is, the less precise the individual-level estimates of the underlying trait. Statistical evidence of DIF may be revealed in large field studies, but the magnitude of detected DIF may be small relative to the level of precision of the testing instrument. Although to the authors' knowledge, this has not been suggested before, perhaps the first test of the practical importance of DIF is one that expresses the magnitude of bias in underlying latent trait estimates as a fraction of the measurement error of the test instrument.

Summary and Conclusion

Recommendations regarding the study of measurement bias can be drawn from various bodies of research. First, as shown in the statistical literature and summarized by Carroll (2003), errors in variables can have a large impact on structural relationships in regression models. The effects of the base rate and unreliability on sensitivity and specificity of measures has been well reviewed in the epidemiological literature (e.g., Dodd & Korn, 2008), and more recently Aguinis et al. (2010) studied the effects of unreliability and other factors on preemployment tests, concluding that there is a need for revival of test bias research. Second, the method most commonly used to estimate reliability and, unfortunately, to compare measures across groups, that is, coefficient alpha (Cronbach, 1951), is inadequate for the task. Factors affecting coefficient alpha and various forms of the intraclass correlation coefficient, and reasons why they should not be compared in cross-cultural studies, have been well discussed (e.g., Feinstein & Cicchetti, 1990; Hambleton, Swaminathan, & Rogers, 1991; Kraemer & Bloch, 1988; Teresi & Holmes, 2001). As reviewed in a recent series of influential articles in the psychometrics literature (e.g., Bentler, 2009; Green & Yang, 2009; Reise et al., 2010; Revelle & Zinberg, 2009; Sijtsma, 2009), a move toward latent variable models and reliability estimates derived from such models is recommended. An important point is that latent variable models are not a panacea; they must be specified carefully. Third, reliability estimates, even those derived from latent variable models, are not properties of a measure and will change according to the groups studied. Thus, even in professionally developed measures, the psychometric performance of the measure, including factorial invariance, needs to be studied when applied to new samples or groups. As reviewed in this chapter, there is a complex relationship between measurement invariance and prediction invariance. It is important to include the study of item-level measurement invariance as a step in the process of

ensuring that measures are equivalent across groups differing on characteristics such as race, gender, age, and ethnicity.

Borsboom (2006a) has discussed why psychometrics has in general failed to solve major issues in psychological measurement and become part of the standard armamentarium of quantitative psychology. Borsboom cited work by Millsap (1997), updated in Millsap (2007a), demonstrating that a measure for which measurement invariance holds will show prediction noninvariance with respect to an external criterion, and when prediction invariance holds, measurement invariance will generally not hold. Prediction invariance does not imply measurement invariance, and bias in measures can lead to selection errors (Millsap, 2007). Despite these results, Borsboom pointed out that the official position of the field as embodied in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) is that that the primacy of evidence for bias is granted to prediction invariance. Hunter and Schmidt (2000), on the basis of an extensive review (Schmidt & Hunter, 1998), concluded that the issue of item bias was resolved and irrelevant for racial and gender bias in ability and achievement tests. However, echoing Borsboom, the view advanced in this chapter is that the issue of item bias is not dead. Rather, there is a need for continued applied examples that are executed and interpreted appropriately and that move beyond aggregate reports of statistical effects (such as DIF statistics) to describe the impact of bias in terms of differences at the level of individual score assignment and selection decisions.

APPENDIX 8.1: ITEM RESPONSE THEORY–BASED DIFFERENTIAL ITEM FUNCTIONING SOFTWARE

SOFTWARE FOR DIMENSIONALITY ASSESSMENT

Polychoric and polyserial correlations can be estimated using structural equation modeling packages such as MPlus (L. K. Muthén & Muthén, 2012). The explained common variance as well as McDonald's omega can be estimated using most software and generated from both exploratory and confirmatory factor analyses. R software (R Development Core Team, 2008) can produce the Schmid–Leiman (1957) solution, and McDonald's omega statistics as well as several others recommended by Revelle and Zinbarg (2009), are contained in the Psych package developed in R (http://www.R-project.org). EQSIRT (E. J. C. Wu & Bentler, 2011) can also be used to calculate most of the measures. Final loadings can be estimated using software such as MPlus (L. K. Muthén & Muthén, 2012) and Item Response Theory (IRT) for Patient-Reported Outcomes (Cai et al., 2012).

SOFTWARE FOR DIFFERENTIAL ITEM FUNCTIONING ANALYSES

In this section we describe some software that can be used for DIF analyses. The field is rapidly expanding, and a complete list is unattainable. This description provides a discussion of some of the more commonly used software approaches in the field at this time.

Logistic and Ordinal Logistic Regression

Logistic regression (Swaminathan & Rogers, 1990) and ordinal logistic regression (Zumbo, 1999) methods can be performed with most standard software. A modification, IRTOLR (Crane et al., 2004), uses estimates from a latent variable IRT model rather than the traditional observed score conditioning variable and incorporates effect sizes into the uniform differential item functioning (DIF) detection procedure. Difwithpar (Crane et al., 2006) allows the user to specify the criteria for DIF, for example, statistical tests of uniform and nonuniform (Swaminathan & Rogers, 1990) and effect size modification based on changes in the pseudo-R^2 in nested models (Zumbo, 1999) or a change in the coefficient criterion for uniform DIF (Crane et al., 2004).

IRTOLR

Lordif (Choi, Gibbons, & Crane, 2011) incorporates several logistic and ordinal logistic regression models that condition on a latent variable. On the basis of the difwithpar framework (Crane et al., 2006), lordif was developed to perform ordinal logistic

regression with an iteratively purified IRT trait estimate as the matching criterion. A Monte Carlo simulation approach permits empirically derived threshold values and effect size and impact measures (Kim, Cohen, Alagoz, & Kim, 2007). Lordif uses ltm (Rizopoulus, 2006, 2009) to obtain IRT item parameter estimates for the graded response model (Samejima, 1969) and the Design package for ordinal logistic regression.

Differential Item Functioning With Rasch Models

The conditional maximum likelihood estimation procedure for extended Rasch modeling (Mair & Hatzinger, 2007a, 2007b), including Wald tests (Glas & Verhelst, 1995) and the Andersen (1973) likelihood ratio tests, can be found in the eRm package in R (R Development Core Team, 2008). The displacement values can be requested by using Winsteps (Linacre, 2005, 2009). Many other software packages can be used to conduct Rasch analysis, for example, Rasch unidimensional measurement models (Andrich, Lyne, Sheridan, & Luo, 2003).

Item Response Theory Log-Likelihood Ratio Modeling

The IRT log-likelihood methods for DIF tests for both the dichotomous and the graded response models are provided by likelihood ratio tests (Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Thissen, Steinberg, & Wainer, 1993) used in IRT Likelihood-Ratio Tests for Differential Item Functioning (www.unc.edu/~dthissen/dl.html) and MULTILOG (Thissen, 1991, 2001).

IRTPRO

IRTPRO (Cai et al., 2009, 2012) includes bifactor and multidimensional item response models and tests of model assumptions. In the context of DIF testing, it incorporates the Wald test (Lord's [1980] chi-square) for testing DIF across multiple groups (see also Langer, 2008) but with more accurate parameter error variance–covariance matrices computed using the supplemented expectation–maximization algorithm (Cai, 2008). Contrasts among several groups can be performed using R. Graphics for the test characteristic curves and information functions are provided. The likelihood ratio test using the IRTLRDIF anchoring approach is incorporated; however, this process is relatively labor intensive and, given the asymptotic equivalence of the logistic regression and Wald tests, probably not efficient. The Wald test uses an anchor-all, test-all method; however, purification steps can be introduced by additional runs selecting anchor sets without DIF. The two-tier full-information maximum marginal likelihood factor analysis model (Cai, 2010), used to test hypotheses relating to DIF in a variety of study designs, is available in IRTPRO (Cai et al., 2012) and MEDPRO (Thissen, 2009).

Differential Functioning of Items and Tests

The noncompensatory DIF effect size measure and the contribution of the item-level DIF to the total test (the differential functioning at the test level index) can be found in differential functioning of items and tests (Oshima, Kushubar, Scott, & Raju, 2009).

Multiple Indicator–Multiple Cause and Multiple-Group Confirmatory Factor Analysis

Parameter estimates for multiple-group confirmatory factor analysis models can be obtained with Mplus software. A robust parameter estimation procedure is based on a mean- and variance-adjusted weighted least squares procedure (B. O. Muthén, du Toit, & Spisic, 1997; L. K. Muthén & Muthén, 2012). Mplus models can be estimated using sampling weights.

EQSIRT

EQSIRT (E. J. C. Wu & Bentler, 2011) includes traditional approaches to structural equation modeling estimation as well as those common to IRT. Similar to the IRTLR, EQSIRT constrains all parameters to be equal and tests the equality constraints for significant DIF. Parameter change is examined using the Lagrange multiplier (C.-P. Chou & Bentler, 1990). Additionally, differences in location parameters between reference and focal groups can be tested in a nonlinear mixed IRT model with a group covariate (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). EQSIRT includes the capacity to evaluate magnitude and impact.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*, 648–680. doi:10.1037/a0018714

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140. doi:10.1007/BF02291180

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). RUMM2020 [Computer software]. Perth, Western Australia, Australia: RUMM Laboratory.

Angel, R. J. (2006). Narrative and the fundamental limitations of quantification in cross-cultural research. *Medical Care, 44*(Suppl. 3), S31–S33. doi:10.1097/01.mlr.0000245428.03255.cf

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–25) Hillsdale, NJ: Erlbaum.

Baker, D. W., Wolf, M. S., Feinglass, J., Thompson, J. A., Gazmararian, J. A., & Huang, J. (2007). Attitudes toward health care providers: Collecting information about patients' race, ethnicity, and language. *Medical Care, 45*, 1034–1042. doi:10.1097/MLR.0b013e318127148f

Benjamini, Y., & Hochberg, Y. (1995). Controlling for the false discovery rate: A practical powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 57*, 289–300.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137–143. doi:10.1007/s11336-008-9100-1

Blank, R. M., Dabady, M., & Citro, C. F. (2004). *Measuring racial discrimination.* Washington, DC: National Academies Press.

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika, 71*, 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D. (2006b). Can we bring about a velvet revolution in psychological measurement? A rejoinder to commentaries. *Psychometrika, 71*, 463–467. doi:10.1007/s11336-006-1502-3

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods, 13*, 75–98. doi:10.1037/1082-989X.13.2.75

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309–329. doi:10.1348/000711007X249603

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581–612. doi:10.1007/s11336-010-9178-0

Cai, L., du Toit, S. H. L., & Thissen, D. (2009). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Cai, L., Thissen, D., & Du Toit, S. H. L. (2012). IRTPRO version 2.1 [Computer software]. Skokie IL: SSI Scientific Software International.

Carroll, R. J. (2003). Variances are not always nuisance parameters. *Biometrics, 59*, 211–220. doi:10.1111/1541-0420.t01-1-00027

Carroll, R. J., Gallo, P., & Gleser, L. J. (1985). Comparison of least squares and errors-in-variable regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association, 80*, 929–932.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomous scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333–353. doi:10.1111/j.1745-3984.1996.tb00496.x

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: A R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulation. *Journal of Statistical Software, 39,* 1–30.

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier and Wald test. *Multivariate Behavioral Research, 25*, 115–136. doi:10.1207/s15327906mbr2501_13

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement, 70*, 717–731. doi:10.1177/0013164410379322

Christensen, H., Jorm, A. F., Mackinnon, A. J., Korten, A. E., Jacomb, P. A., Henderson, A. S., & Rodgers, B. (1999). Age differences in depression and anxiety symptoms: A structural equation modeling

analysis of data from a general population sample. *Psychological Medicine, 29,* 325–339. doi:10.1017/S0033291798008150

Clark, L. A. (2006). When a psychometric advance falls in the forest. *Psychometrika, 71*, 447–450. doi:10.1007/s11336-006-1500-5

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335–350. doi:10.1177/014662169301700402

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*, 15–26. doi:10.1177/014662169602000102

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P., Cohen, J., Teresi, J., Marchi, P., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement, 14*, 183–196. doi:10.1177/014662169001400207

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care*, *44*(Suppl. 3), S115–S123. doi:10.1097/01.mlr.0000245183.28384.ed

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241–256. doi:10.1002/sim.1713

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555

Dodd, L. E., & Korn, E. I. (2008). Lack of generalizability of sensitivity and specificity of treatment effects. *Statistics in Medicine, 27*, 1734–1744. doi:10.1002/sim.3101

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368. doi:10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the Mini-Mental State Examination: An application of the Mantel-Haenszel and standardization procedures. *Medical Care, 44*(Suppl. 3), S107–S114. doi:10.1097/01.mlr.0000245182.36914.4a

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale, NJ: Erlbaum.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543–549. doi:10.1016/0895-4356(90)90158-L

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York, NY: Wiley.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*, 309–326. doi:10.1177/01466219922031437

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 49*, 251–264.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored. An illustration with the Center for Epidemiological Studies Depression Scale. *Educational and Psychological Measurement, 63*, 65–74. doi:10.1177/0013164402239317

Glas, C. A. W., & Verhelst, N. (1995). Testing the Rasch model. In G. H. Fisher & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 69–96). New York, NY: Springer.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12*, 133–157. doi:10.1177/002188637601200201

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121–135. doi:10.1007/s11336-008-9098-4

Gregorich, S. E. (2006). Do self report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(Suppl. 3), S78–S94. doi:10.1097/01.mlr.0000245454.12228.8f

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence instrument applicable on adolescents? A latent trait analysis using Rasch modeling. *Personality and Individual Differences, 36*, 955–968. doi:10.1016/S0191-8869(03)00164-8

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(Suppl. 3), S182–S188. doi:10.1097/01.mlr.0000245443.86671.c4

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement, 71*, 306–324. doi:10.1177/0013164410384856

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6*, 151–158. doi:10.1037/1076-8971.6.1.151

Jennrich, R. I., Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76,* 537–549. doi:10.1007/s11336-011-9218-4

Jennrich, R. I., Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77,* 442–454. doi:10.1007/s11336-012-9269-1

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329–349. doi:10.1207/S15324818AME1404_2

Johnson, T. P. (2006). Methods and frameworks for cross-cultural measurement. *Medical Care, 44*(Suppl. 3), S17–S20. doi:10.1097/01.mlr.0000245424.16482.f1

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *Medical Care, 44*(Suppl. 3), S124–S133. doi:10.1097/01.mlr.0000245250.50114.0f

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the Mini-Mental State Examination: Effects of differential item functioning. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 57*, 548–P558. doi:10.1093/geronb/57.6.P548

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36,* 347–387. doi:10.1207/S15327906347-387

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345–355. doi:10.1177/014662169802200403

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93–116. doi:10.1111/j.1745-3984.2007.00029.x

Kraemer, H. C., & Bloch, D. A. (1988). Kappa coefficients in epidemiology: An appraisal of a reappraisal. *Journal of Clinical Epidemiology, 41*, 959–968. doi:10.1016/0895-4356(88)90032-7

Krause, N. (2006). The use of qualitative methods to improve quantitative measure of health-related constructs. Measurement in a multi-ethnic society. *Medical Care, 44*(Suppl. 3), S34–S38. doi:10.1097/01.mlr.0000245429.98384.23

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model—Some critical suggestions on traditional approaches. *International Journal of Testing, 5*, 377–394. doi:10.1207/s15327574ijt0504_3

Langer, M. M. (2008). *A re-examination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation, University of North Carolina at Chapel Hill). Retrieved from http://search.lib.unc.edu/search?R=UNCb5878458

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusion based on observed test score data. *Psicológica, 30*, 343–370.

Linacre, J. M. (2005–2009). WINSTEPS Rasch Measurement software for persons and items [Computer software]. Chicago, IL: Mesa Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (with A. Birnbaum). (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Ma, Y., Hart, J. D., Janicki, R., & Carroll, R. J. (2011). Local and omnibus goodness-of-fit tests in classical measurement error models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *73,* 81–98. doi:10.1111/j.1467-9868.2010.00751.x

Mair, P., & Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26–43.

Mair, P., & Hatzinger, R. (2007b). Extended Rasch modeling: The R package Rm for the application of IRT models in R. *Journal of Statistical Software, 20*, 1–20.

Manly, J. J. (2006). Deconstructing race and ethnicity: Implications for measurement of health outcomes. *Medical Care, 44*(Suppl. 3), S10–S16. doi:10.1097/01.mlr.0000245427.22788.be

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*, 344–362. doi:10.1037/1082-989X.11.4.344

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects

of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*, 357–367. doi:10.1177/014662169802200404

McArdle, J. J., Fisher, G. G., & Kadlec, K. M. (2007). Latent variable analyses of age trends of cognition in the Health and Retirement Study 1992–2004. *Psychology and Aging, 22*, 525–545. doi:10.1037/0882-7974.22.3.525

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika, 76*, 511–538. doi:10.1007/s11336-011-9223-7

McHorney, C. A., & Fleishman, J. A. (2006). Assessing and understanding measurement equivalence in health outcome measures. *Medical Care, 44*(Suppl. 3), S205–S210. doi:10.1097/01.mlr.0000245451.67862.57

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307. doi:10.1037/0033-2909.115.2.300

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177–185. doi:10.1007/BF02289699

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. doi:10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(Suppl. 3), S69–S77. doi:10.1097/01.mlr.0000245438.73837.89

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single factor case. *Psychological Methods, 2*, 248–260. doi:10.1037/1082-989X.2.3.248

Millsap, R. E. (2007a). Invariance in measurement and prediction revisited. *Psychometrika, 72*, 461–473. doi:10.1007/s11336-007-9039-7

Millsap, R. E. (2007b). Structural equation modeling made difficult. *Personality and Individual Differences, 42,* 875–881. doi:10.1016/j.paid.2006.09.021

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334. doi:10.1177/014662169301700401

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389–402. doi:10.1177/014662169201600411

Millsap, R. E., & Meredith, W. (1994). Statistical evidence in salary discrimination studies: Nonparametric inferential conditions. *Multivariate Behavioral Research, 29,* 339–364. doi:10.1207/s15327906mbr2904_2

Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika, 51*, 177–195. doi:10.1007/BF02293979

Morales, L. S., Flowers, C., Gutierrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the differential item and test functioning (DFIT) framework. *Medical Care, 44*(Suppl. 3), S143–S151. doi:10.1097/01.mlr.0000245141.70946.29

Mui, A. C., Burnette, D., & Chen, L. M. (2001). Cross-cultural assessment of geriatric depression: A review of the CES-D and GDS. *Journal of Mental Health and Aging, 7,* 137–164.

Murphy, J. M., Laird, N. M., Monson, R. R., Sobol, A. M., & Leighton, A. H. (2000). A 40-year perspective on the prevalence of depression: The Stirling County Study. *Archives of General Psychiatry, 57*, 209–215. doi:10.1001/archpsyc.57.3.209

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115–132. doi:10.1007/BF02294210

Muthén, B. (1989a). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Muthén, B. (1989b). Using item-specific instructional information in achievement modeling. *Psychometrika, 54,* 385–396. doi:10.1007/BF02294624

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81–117. doi:10.2333/bhmk.29.81

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Los Angeles: University of California.

Muthén, B. O., du Toit, S. H., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Los Angeles: University of California, Los Angeles.

Muthén, B. O., & Hsu, J.-W. H. (1993). Selection and predictive validity with latent variable structures. *British Journal of Mathematical and Statistical Psychology, 46*, 255–271. doi:10.1111/j.2044-8317.1993.tb01015.x

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133–142. doi:10.2307/1164840

Muthén, L. K., & Muthén, B. O. (2012). *MPLUS Users Guide* (6th ed.). Los Angeles, CA: Author.

Nápoles-Springer, A. M., Santoyo, J., O'Brien, H., & Stewart, A. L. (2006). Using cognitive interviews to develop surveys in diverse populations. *Medical Care, 44*(Suppl. 3), S21–S30. doi:10.1097/01.mlr.0000245425.65905.1d

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S – X^2: An item fit index for

use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298. doi:10.1177/0146621603027004004

Orlando Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental State Examination. *Medical Care, 44*(Suppl. 3), S134–S142. doi:10.1097/01.mlr.0000245251.83359.8c

Oshima, T. C., Kushubar, S., Scott, J. C., & Raju, N. S. (2009). *DFIT for Windows user's manual: Differential functioning of items and tests.* St. Paul, MN: Assessment Systems.

Pérez-Stable, E. J. (2007). Language access and Latino health care disparities. *Medical Care, 45,* 1009–1011. doi:10.1097/MLR.0b013e31815b9440

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23–37. doi:10.1177/014662169501900104

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definitions and operation criteria. *Statistics in Medicine, 8*, 431–440. doi:10.1002/sim.4780080407

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502. doi:10.1007/BF02294403

Raju, N. S. (1999). DFITP5: A Fortran program for calculating dichotomous DIF/DTF [Computer program]. Chicago: Illinois Institute of Technology.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, E. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33*, 133–147. doi:10.1177/0146621608319514

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353–368. doi:10.1177/014662169501900405

Ramirez, M., Teresi, J., Holmes, D., Gurland, B., & Lantigua, R. (2006). Differential item functioning and the Mini Mental State Examination (MMSE): Overview, sample and issues of translation. *Medical Care, 44*(Suppl. 3), S95–S106. doi:10.1097/01.mlr.0000245181.96133.db

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute of Educational Research.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P., Cook, K. F., & Moore, T. M. (in press). A direct modeling approach for evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. Reise & D. Revicki (Eds.), *Handbook of applied item response theory in typical performance assessment*. New York, NY: Taylor & Francis.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544–559. doi:10.1080/00223891.2010.496477

Reise, S. P., Moore, T. M., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement, 71,* 684–711. doi:10.1177/0013164410378690

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measure. *Quality of Life Research, 16*, 19–31.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566. doi:10.1037/0033-2909.114.3.552

Revelle, W., & Zinbarg, R. E. (2009). Coefficient alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74*, 145–154. doi:10.1007/s11336-008-9102-z

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185–205. doi:10.1037/1082-989X.8.2.185

Rizopoulus, D. (2006). Ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*, 1–25.

Rizopoulus, D. (2009). *Ltm: Latent trait models under IRT*. Retrieved from http://cran.rproject.org/web/packages/ltm/index.html

Rogler, L. H. (1989). The meaning of culturally sensitive research in mental health. *American Journal of Psychiatry, 146,* 296–303.

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355–371. doi:10.1177/014662169602000404

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement, 64,* 588–599. doi:10.1177/0013164403261051

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by sex in employment-oriented personality

measures. *Journal of Applied Psychology, 87*, 667–674. doi:10.1037/0021-9010.87.4.667

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences in cognitive tests. *American Psychologist, 59*, 7–13. doi:10.1037/0003-066X.59.1.7

Sackett, P. R., Schmitt, N., Ellington, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318. doi:10.1037/0003-066X.56.4.302

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society.

Saulny, S. (2011, February 10). Race remixed: In a multiracial nation, many ways to tally. *New York Times.*

Schmid, L., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53–61. doi:10.1007/BF02289209

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of over 85 years of research findings. *Psychological Bulletin, 124,* 262–274. doi:10.1037/0033-2909.124.2.262

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. doi:10.1007/s11336-008-9101-0

Snow, T. K., & Oshima, T. C. (2009). A comparison of unidimensional and three-dimensional differential item functioning analysis using two-dimensional data. *Educational and Psychological Measurement, 69,* 732–747. doi:10.1177/0013164409332223

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497–508. doi:10.1037/0021-9010.89.3.497

Steele, C. M., & Aronson, J. (1995). Stereotypic threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811. doi:10.1037/0022-3514.69.5.797

Steele, C. M., & Aronson, J. A. (2004). Stereotype threat does not live by Steele and Aronson (1995) alone. *American Psychologist, 59,* 47–48. doi:10.1037/0003-066X.59.1.47

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589–617. doi:10.1007/BF02294821

Stout, W. F., & Roussos, L. A. (1995). *SIBTEST users manual.* Urbana–Champaign: University of Illinois, Department of Statistics.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408. doi:10.1007/BF02294363

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*(Suppl. 3), S152–S170. doi:10.1097/01.mlr.0000245142.74628.ab

Teresi, J. A., & Holmes, D. (2001). Some methodological guidelines for cross-cultural comparisons. *Journal of Mental Health and Aging, 7*, 13–19.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651–1683. doi:10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AID-SIM453>3.0.CO;2-H

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E., . . . Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measure physical functioning ability and general distress. *Quality of Life Research, 16*, 43–68. doi:10.1007/s11136-007-9186-4

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. E., Crane, P. K., Jones, R. N., . . . Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly, 51,* 148–180.

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of Aging and Health, 24,* 1044–1076. doi:10.1177/0898264312436877

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of differential item functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly, 50*, 538–612.

Teresi, J. A., Stewart, A., Morales, L., & Stahl, S. (2006). Measurement in a multi-ethnic society: Overview to the special issue. *Medical Care, 44*(Suppl. 3), S3–S4. doi:10.1097/01.mlr.0000245437.46695.4a

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. Retrieved from http://www.unc.edu/~dthissen/dl.html

Thissen, D. (2009). The MEDPRO project: An SBIR project for a comprehensive IRT and CAT software system—IRT software. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved from http://psych.umn.edu/psylabs/CATCentral

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118–128. doi:10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false discovery rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77–83. doi:10.3102/10769986027001077

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. doi:10.1177/109442810031002

Wainer, H. (1993). Model-based standardization measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Erlbaum.

Wang, W.-C., & Shih, C.-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34*, 166–180. doi:10.1177/0146621609355279

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42–57. doi:10.1177/0146621607314044

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison of two group analysis. *Multivariate Behavioral Research, 44*, 1–27. doi:10.1080/00273170802620121

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH tests and IRTLRDIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement, 35*, 145–164. doi:10.1177/0146621610377450

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*, 339–361. doi:10.1177/0146621611405984

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation, 12*, 1–26.

Wu, E. J. C., & Bentler, P. M. (2011). *EQS-IRT—A user-friendly IRT program*. Encino, CA: Multivariate Software.

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summed scores in predicting outcomes. *Educational and Psychological Measurement, 72*, 453–463. doi:10.1177/0013164411419846

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

CHAPTER 9

TEST DEVELOPMENT STRATEGIES

Neal M. Kingston, Sylvia T. Scheuring, and Laura B. Kramer

Strategy is an often misused term. Strategy is a plan of action designed to achieve high-level goals. Strategy focuses on the big picture and the long term. Good strategies seldom change. *Tactics,* however, are ways of accomplishing strategy. Tactics change with circumstance. In business or the military, tactics are designed to seek out weaknesses in an ever-changing landscape. In research and development, tactics should change to take advantage of new methods or new market forces. This chapter focuses on test development strategy but also considers high-level tactics that might be used to implement the chosen strategies. Other chapters provide detail about test development tactics (e.g., Chapters 13, 14, 16, 18, and 20, this volume).

With this definition in mind, the goals of test development are usually to create a measure that best supports intended inferences users will make from test scores. Those intended inferences will vary, and thus strategies will vary. A test may be designed to select people for a limited number of positions, as is often the case with college admissions tests or employment tests. A different test might be designed to provide feedback that improves instruction. A third test might be needed to support inferences regarding optimal treatment for a person who has a mental illness. The test development strategies for each of these tests would be different.

There are many types of tests, including ability, achievement, interest, and personality tests. Any of these types can exist as a single measure or as a battery of related tests. An example of some test development strategies (or an element of such strategies) is facilitating the comparison of examinees (a) with others (norm referenced), (b) against a standard (domain referenced, criterion referenced, or standards based), or (c) with other attributes they possess (ipsative).

In addition to the form of the comparison, score interpretation can be based on what an examinee can do at the present, or it can be a prediction of what the examinee will do in the future. Test development strategies for each type of test have many similarities but also important differences. This chapter focuses on the most common strategic approaches in the test development enterprise but does not cover all possibilities.

Thoughtful test development will begin with the end in mind. What questions should be answered with the test results? What are the desired outcomes of the test? Will the test be used to change practices, gain information, or implement rewards or consequences? What inferences are anticipated to be drawn from the test results? Will the scores be used for diagnosis, to rank order examinees, or to sort them into categories? What is the nature of the reports that need to be provided? Will the data be reported for an individual, disaggregated into subgroups, or reported at the population level only? Will a single score be extracted, or are there multiple facets on the test from which information must be derived? If these issues are not faced from the beginning, it could result in a test that does not deliver what is needed.

Regardless of the type of test being created, an effective test development strategy begins with a

DOI: 10.1037/14047-009
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

coherent test design. Once there is a well-designed assessment, test development is a matter of developing test items (or selecting items from an existing pool) according to the blueprint and schedule. In addition to item writing, sufficient time must be allocated to item editing as well as reviews for construct clarity, accuracy, age appropriateness, consistency of style (content review), and appropriately balanced cultural, socioeconomic, and gender perspectives (bias review). Some reporting designs may require that time be set aside for standard-setting procedures. Once items have been reviewed and developed, piloting, field testing, and final forms selection are performed, culminating in a test in which the content supports the intended inferences (validity) and psychometric reliability.

STRATEGIC APPROACH TO TEST DESIGN

Coherent test design requires incorporation of information across multiple disciplines, including psychometrics, technology, content development, finance, and project management. The greatest errors in test design are usually the result of ignoring information from one or more of these major disciplines. In addition to considering information across these disciplines, test design incorporates an understanding of the evolution of the assessment over time.

Well before item writing or item selection from an existing pool of items begins, a test design should be developed. A good test design addresses the following:

- a definition of the construct to measure;
- a detailed description of the target population;
- gender balance;
- socioeconomic status distribution;
- educational background;
- reporting requirements;
- description of the audience for each report;
- scores and subscores that will be provided (e.g., raw scores, scale scores, percentage correct, percentiles);
- interpretive information to be provided to each audience to meet the measurement goals;
- relative or absolute timeline of availability for each report (e.g., immediately, after 6 weeks of operation, after 1 year of operation, after 1,000 test takers have taken the test);
- measurement domain for each report, if it not the same as the measurement domain for the overall assessment (e.g., whether group reports will provide information on subdomains or change scores not reported on individual reports);
- population range for each report (individual, subpopulation, or total population);
- administration requirements (e.g., such as individual vs. group, time limits, and stopping rules);
- test delivery options to be made available (e.g., online, paper and pencil);
- allowable tools or manipulatives (e.g., rulers, calculators, spell check, highlighters, graph paper, or scratch paper);
- accommodations to be supported;
- underlying psychometric model;
- item ordering (linear, testlet, item by item, self-adaptive, matrix, or mixed models);
- content range, number, and type of each item to be included on the test (e.g., multiple choice, short answer, essay, performance tasks);
- item-scoring approach (e.g., answer key, latent semantic analysis, human application of rubrics);
- item-score aggregation approach (e.g., number correct, formula scoring, item response theory [IRT] pattern scoring, objective performance index);
- scaling (e.g., setting means and standard deviations, area transformations);
- statistical methods that will be used to evaluate students' responses;
- a blueprint for test construction;
- the number and organization of test sessions (sections within the test);
- the percentage of coverage for each category to be tested;
- a detailed description of the part of each category to be measured and the type of items to be used for each category (i.e., item specifications);
- any constraints to be applied to shared stimuli (material provided to the test taker that is used to perform more than one task);
- documentation of the expected evolution of the test over time;
- a schedule for pilot testing, field testing of items, and norming of forms (if required);

- considerations in item or form replacement because of item exposure limits or legal rulings on item release (e.g., sunshine laws);
- a proposed strategy for improving or replacing items and forms over time (e.g., stand-alone [explicit] or embedded field testing); and
- a plan for equating test forms over time.

The more thoughtful and detailed this test design, the more likely it is that the resulting test will meet its measurement goals.

Creating the Test Design

Test design is all about the validity of the inferences a test is intended to support. Every design decision a test developer makes should facilitate users' making appropriate inferences about individuals, groups, or programs being evaluated on the (at least partial) basis of test scores. The keys to an effective test design are to know what questions should be asked and how to interpret the answers. These questions will relate to inferences directly, such as whether a test will be used for prediction. Other issues, such as administration conditions, will have a somewhat indirect effect on inferences. Not all questions are pertinent to all types of tests—for example, although some issues are common to the design of personality and academic achievement tests, some are not. However, the answers to all pertinent questions will affect the quality of the inferences test users will make on the basis of test scores.

The questions that follow are organized around four categories. Question 1 concerns the purpose of the test: predictive or descriptive? Questions 2–7 concern score reporting. Although reporting is the end of the testing process, considering it early helps to clarify many decisions that need to be made. Questions 8 and 9 concern test administration. Administration can place constraints on a testing program that need to be considered early. Questions 10 and 11 consider the social and political environment in which testing occurs.

1. Will the test be used to predict or to describe?
2. How many scores will the test support?
3. Will the reports compare test takers' performance with each others', with a specific set of goals (standards), or with their own?
4. Will scores be reported for individual examinees, groups, or both?
5. Will the test be used to report growth or change in the examinee's performance over time?
6. Who will get reports? Which reports will they receive? How will performance data be categorized for each audience?
7. How soon will reports need to be made available after testing is completed?
8. Should this testing program be administered on computer or on paper?
9. Should examinees have limited time to take the test?
10. What are the stakes? How likely is cheating?
11. How transparent should the testing program be?

In the following sections, we address how different answers to these questions will shape the test design. We also discuss how to determine when answers may indicate conflicting measurement or development goals. When a given test design attempts to meet too many conflicting goals, the validity of the resulting information is likely to suffer. In such cases, it may be better to design two or more different tests, rather than a single one, to reach all measurement goals. It is also useful to know up front whether the measurement goals and budget or timeline are in direct conflict. Some insight is provided here as well.

Will the Test Be Used to Predict or to Describe?

If a test will be used to predict the performance of test takers on a future criterion, there are two primary strategies: Build a test to measure (a) general characteristics that have been demonstrated to be predictive of future performance or (b) current achievement in the domain of interest. The former approach was used in the United States for the development of academic and personnel selection tests throughout much of the 20th century and is still popular. Stemming from early work on intelligence testing, aptitude tests have been developed to have high loadings on general intelligence and specific verbal and mathematical abilities using item types such as analogies, antonyms, and pattern matching. People who possess high levels of these

aptitudes tend to learn new information faster and more readily than those who do not. Examples of selection tests that have been used to predict future performance on the basis of aptitudes include the pre-1994 SAT,[1] the GRE General Test, and the Law School Admission Test.

Other personal characteristics relate to academic success, such as perseverance. These characteristics could also be used as the strategic basis for the development of predictive selection tests. Unfortunately, the state of the art has not yet advanced to the point at which such measures are as near as reliable or predictive as tests of aptitude. The Educational Testing Service has been a leading research organization in this arena, with work on a strivers index in the 1980s and 1990s and the current graduate school Personal Potential Index.

Despite relatively low academic aptitude (as measured by such tests), some students learn well by depending on either unmeasured cognitive strengths or the strength of noncognitive factors (such as working harder). An alternative to trying to measure noncognitive skills directly is to look at current achievement, because this achievement has likely benefited from these noncognitive attributes. The ACT has used this approach, focusing primarily on general reading and the reading of science and social studies materials. The post-2005 SAT Critical Reading component uses a similar strategy. The GRE Subject Tests have always used this strategy, and in most programs for which there exists an appropriate Subject Test (currently, Biochemistry, Biology, Chemistry, Computer Science, Literature in English, Mathematics, Physics, and Psychology), they have been shown to be more predictive than the General Test. Furthermore, in about half of all programs, GRE Subject Tests have also been shown to be more predictive than undergraduate grade point average (Schneider & Briel, 1990).

Some tests contain elements of both strategies. Interest inventories describe an examinee's current interests, but this description can also be connected to predictions of future job satisfaction.

How Many Scores Will the Test Report?

Some test developers intend for their tests to measure a single construct and report a single score. Examples of such tests include the GRE Literature in English examination and the Apgar test used to assess the health of newborn children. Other tests, such as the revised NEO Personality Inventory, the Stanford-Binet 5, and the ACT, provide multiple scores. A test developer will want to report a single score when there is a narrowly defined construct or no theoretical basis or practical utility for multiple scores. For many constructs, however, there is either a well-supported theoretical basis for subscores, such as the Cattell–Horn–Carroll theory of intelligence (Carroll, 1993) or the Big Five personality model (Digman, 1990), or empirical evidence in support of diagnostic utility, such as for the 16 Personality Factor Questionnaire (Dana, Bolton, & Gritzmacher, 1983).

True diagnostic assessment must provide information that supports differentiated prescription. The burden of proof rests with the test publisher to show that an appropriate differentiated prescription leads to improved outcomes, whether those outcomes are academic learning, job performance, mental health, or other.

Broad content categories are not usually usefully diagnostic. Knowing an eighth-grade mathematics student is weak on word problems without knowing whether the issue is reading, extracting key information, setting up the problems, specific mathematical methods, or computation leaves a teacher with too little information to efficiently plan an instructional intervention. Unreliable scores, based on too few items, are also likely to lead to incorrect inferences and subsequent actions. The interested reader may want to look at the work of Sinharay, Haberman, and Puhan (2007) for a proposed statistical approach to determining whether subscores might be useful. Some possible strategies for creating useful subscores follow, but first a note related to reliability.

Reliability has to do with the consistency of test scores taken on multiple occasions or with alternative test editions (see Chapters 2 and 3, this volume).

[1]Beginning with the removal of antonym items in 1994 and culminating with the removal of analogy items and the renaming of the verbal measure as Critical Reading in 2005, the College Board has changed this aspect of its test development strategy.

If an individual took a test many times and each time obtained a very different score (i.e., the test had low reliability), this inconsistency would lead to inconsistent inferences (and thus low validity). The question remains, How low can reliability be before a subscore is no longer useful? To a large extent, the answer depends on the consequences of the decisions based in whole or in part on the test scores. A test used to determine whether people can practice their chosen profession (e.g., to be licensed as a physician or a lawyer) has severe economic consequences. Alternatively, a test that merely provides advice as to what career paths might be of interest or an interim assessment that leads to recommendations about what a student most needs to study has a lower degree of consequence.

Because subscores are relatively short, they are often significantly less reliable than typical total scores, and test developers must often think about ways to increase the reliability of subscores. A discussion of some ways of increasing subscore reliability follows.

Increase test length. Consider a diagnostic test designed to measure test-taker knowledge about subtraction. The test can contain many items that investigate particular aspects of subtraction, such as single digit, double digit without borrowing, and double digit with borrowing. How many items are sufficient for traditional subscores? One would need to look at reliability data to determine the number of items. However, if no significant decisions about individuals will be made on the basis of the subscores and if both the items were written narrowly enough and the scores were based on summing the number of dichotomous (right–wrong) responses, perhaps as few as six to 10 items would do. Other circumstances might require subscores be based on 40 or more items.

Write distractors to support diagnoses. Consider the previously mentioned diagnostic test designed to measure test-taker knowledge about subtraction. One can write incorrect answer choices that include the answer an examinee would get if he or she did not borrow correctly. If a test item is "36 −17 = ___," examinees who answer 21 or 29 are both incorrect but are showing different misunderstandings of subtraction. In the first case, the student consistently subtracted the smaller number from the larger number. In the second, the student did not borrow correctly. This approach can (and should) be combined with other approaches.

Include items in more than one score. If a test has 60 items and 12 content categories, resulting subscores may be insufficiently reliable. However, sometimes items tap into more than a single aspect of content. Alternatively, as is common in science assessments, each item may tap into content and process dimensions. Some test developers try to overcome this problem by counting each item in more than one subscore. If each item counted toward two subscores, then each subscore would be based on 10 items. On analysis, this approach would likely result in higher estimated reliability.

Such an approach has a significant drawback. It artifactually increases the correlations among the subscores, because for each examinee item error variances are added to multiple scores (the source of subscore correlations should only be the true scores). Artifactually correlated subscores make it unlikely that a differentiated prescription will be effective. Consider the extreme example in which all subscores for an examinee are the same (and thus there is a correlation of 1.0 across examinees). The resulting subscore information would not be useful.

Use score augmentation methods. Several methods have been developed to borrow statistical strength from items not included in the subscore of interest (e.g., Wainer et al., 2001). Such methods are statistically more sophisticated than directly counting items in more than one score but similarly increase the correlation between subscores.

Explore the use of diagnostic psychometric models. Diagnostic psychometric models (e.g., Rupp, Templin, & Henson, 2010) hold great promise for extracting diagnostic information from well-designed items. Unfortunately, assessment developers do not yet have much practical experience with tests developed using such models.

Investigate the use of models that adapt on the basis of potential diagnostic information rather than item difficulty. Traditional adaptive testing is based on the estimated difficulty of a unidimensional

set of items. Diagnostic scores are most needed when the assumption of unidimensionality is violated such that two examinees with the same total score have different instructional needs. Regardless of the use of a diagnostic psychometric model, choosing items to be administered on the basis of diagnostic considerations is possible, although such methods have not yet been developed.

Will Reports Compare Test Takers' Performances With Each Others', to a Specific Set of Goals (Standards), or With Their Own?

Test scores accrue meaning as users develop more experience making inferences from those scores. This process can be accelerated by providing one or more forms of reference for those scores, which can be done in three common ways. Perhaps the most common approach is norm referencing, in which examinees are rank ordered and compared with each other (see Chapter 11, this volume). Another approach that is particularly common in both state educational testing programs and licensure and certification testing is criterion referencing. Criterion-referenced tests compare each examinee against some criterion external to the test itself. Criterion-referenced tests typically have one or more cut scores that divide examinees into different performance categories—at the very least into those who pass and those who fail. The third type of comparison, ipsative, is particularly common in interest measurement and certain kinds of personality assessments. Ipsative scores rank the strength of traits within an individual. Thus, if two individuals possess the same rankings, even if one is consistently stronger than the other in all measured traits, their scores would be the same.

Each of these approaches requires a different distribution of item difficulties and other item statistics, and thus a different approach to the blueprint.

Normative. If the test will be used to compare test takers' performance with each others', the test will be a normed reference test (NRT). So how does a need for NRT reporting affect the test design? It depends.

To maximally differentiate students when no point of the score distribution is more important than any other, all items can be of middle difficulty. A middle-difficulty item is one for which half the population of interest knows the answer. If multiple-choice items were used, and examinees who did not know the correct answer guessed at random, then middle difficulty can be calculated as

$$p_{md} = 0.5 + \frac{0.5}{n_c},$$

where n_c is the number of answer choices.

Thus, with four-choice items, if half of the examinees know the correct answer, one can assume that 62.5% will answer correctly because one fourth of those who do not know the answer will guess correctly. An alternative approach to designing tests to maximize the efficacy of normative interpretations uses IRT (see Chapter 6, this volume).

Many tests are used to make important decisions in score ranges when there are not many students—for example, tests used as part of the selection process for highly competitive jobs or for placement in special education programs. Tests are sometimes described as having a "ceiling" or a "floor." A ceiling effect occurs when there is not a sufficient number of difficult items to differentiate among very able examinees. Rather than being distributed along a bell-shaped curve, scores pile up at the high end. Similarly, a floor effect occurs when there are too few easy items to differentiate among low-proficiency examinees. Although a test composed of middle-difficulty items will do a good job of differentiating among the majority of examinees, it will do little good for decisions made at the extremes of the score scale. In such cases, it is important to make sure that the test covers content over a broad range of difficulties so that the test score scale has enough ceiling or floor to differentiate among the examinees for whom decisions are being made. If such test takers got most of the items correct or most of the items incorrect, then one would not be able to effectively distinguish them from each other.

Whether one wants to maximally spread out examinees or measure well at several places in the score scale, IRT (Lord, 1980), assuming its assumptions are met, can provide a more precise analysis than can classical test theory. Item information functions and conditional standard errors of

measurement allow one to ascertain measurement precision at each point of the score scale. Therefore, tests can be built that maximize precision where it is needed most.

The most common IRT methods do not always fit data that violate the assumption of local independence, perhaps because sets of items are based on a common stimulus material. Steinberg and Thissen (1996) demonstrated that testlet response theory (Wainer & Kiely, 1987) could be used to improve the measurement of a short Recent Victimization Scale.

In addition to the impact on the blueprint, the choice of norm-referenced interpretations might also have an impact on reporting timelines; it is best to delay reporting of test scores until a sufficiently representative population of test takers has taken the test. If immediate reporting is required on the day the test is made operational, then a field test and norming will need to be done with a representative population before the test is made operational.

There are additional questions the test developer should ask. For the population to be measured, what will the distribution of test-taker performance likely be? Are test takers expected to be mostly high performers, low performers, or normally distributed?

If the test will be used to select the most able examinees (e.g., for scholarships or jobs for which the applicant pool is large), there will need to be enough difficult content to distinguish among these high performers. Similarly, if the test needs to differentiate among low performers, there will need to be more easy items. In other words, wherever the most precise comparisons are needed is where the most items are required.

Criterion referenced. If the test will report proficiency on a set of standards, then the test will be a criterion-referenced test (CRT). Again, the primary development impact here will be on the blueprint. The number of items on a CRT is very dependent on the type of reporting needed.

First, when creating a CRT, one needs to determine how many cut points one wants to have (see Volume 3, Chapter 22, this handbook, for more detail on standard setting). Will a single standard (such as pass–fail) be set, or will multiple categories of proficiency for test takers (e.g., basic, proficient, advanced) need to be reported? If only a single proficiency level needs to be reported, then most of the items on the assessment should be designed to test knowledge and skills near the proficiency level (just below, on, and just above). If test takers need to be classified into multiple proficiency levels, items will need to maximally discriminate near each of the levels that will be reported, keeping in mind that the more items near a proficiency level determination, the more reliable the measurement of that level.

If setting multiple cut scores, it is important to have not only good measurement precision around the cut points, but also a sufficient number of items in general so that interpretations of the scores make sense. If there are five levels of proficiency and a 20-item test, even with fairly high levels of measurement precision, interpreting the difference in performance between adjacent levels can be difficult and misleading. Additionally, the test must have enough items so that none of the cut points fall within the chance range. If the previous hypothetical 20-item test was a four-option multiple-choice test, examinees would be expected to get five items correct just by chance. Thus, the lowest cut point would need to be at least six items to avoid having someone be able to meet a standard that indicates a level of proficiency merely by guessing at random or by answering C to all 20 questions.

Norm referenced and criterion referenced. It is possible, and increasingly common, to provide criterion-referenced and norm-referenced interpretations for a single test form, although it is difficult to do both well because the strategies to support each sometimes conflict. When attempting to provide both types of score referencing, one might decide to make the test significantly longer than would otherwise have been the case and thus have items that address each goal. When combining these reporting types, not all items need to report to both scales. If there is a high-performing group of test takers that needs to be differentiated in the NRT report, it may be necessary to include more difficult items than would be necessary for a CRT assessment. It may also be possible (and potentially

beneficial for cost purposes) to embed an existing NRT test, which has already been normed for the population, into the CRT assessment (although some experts have questioned the appropriateness of the norms in such cases), which is called an *augmented NRT.* By doing this, NRT information can be reported without the expense and time required to separate norming studies. However, there may be a mismatch between the NRT being used and the CRT standards. If that is the case, it may be beneficial to exclude some of the items used for the NRT reporting from the CRT reports. Note that if the mismatched items are removed from the NRT portion of the test entirely, then the same comparisons can no longer be made, nor can the same interpretations be drawn from the complete NRT. The interested reader is referred to Chapter 11 in this volume for more information on the construction of norms.

Ipsative. In many cases, test developers prefer intraindividual comparisons. When people have to make life choices, they must often depend on their greatest strengths or interests, regardless of how those strengths or interests compare with those of others. For example, a person who prefers working with people more than with numbers may feel better off pursuing a career working with people, regardless of whether there are others who prefer working with people even more strongly.

Many tests have been developed to require ipsative interpretations as a by-product of certain test development challenges rather than as an inherent preference for intraindividual comparison. Specifically, social desirability can be a problem in measures of interest and personality. In the attempt to minimize such challenges, some test developers have chosen a forced-choice format (between choices matched on social desirability). Evidence has suggested that this forced-choice approach does not effectively deal with issues such as social desirability, perhaps only hiding them (Baron, 1996). Moreover, Baron (1996) also synthesized the literature that showed that a forced-choice format causes artifactual negative correlations among items, which in turn makes traditional forms of psychometric analysis suspect.

Will Scores Be Reported for Individual Examinees, Groups, or Both?

Who gets reports will have a significant impact on the potential coverage of the testing domain in the blueprint. Some approaches are effective only when reporting group performance. The interested reader is referred to Volume 3, Chapter 23, this handbook for a detailed description of score reporting considerations.

Design considerations when reporting individual performance. One important consideration for testing programs reporting scores of individual students is score comparability. Score comparability is facilitated when everyone takes exactly the same items in the same order. Also, such an approach is less expensive than having multiple forms. With multiple forms, more items need to be developed, and additional analyses, especially equating, need to be performed. Unfortunately, when individual scores are reported, examinees are sometimes tempted to look at their colleagues' responses. One strategy test developers may choose to use is the deployment of multiple forms at each administration.

The multiple form strategy can be implemented in a variety of ways. The simplest version is to create forms built to the same test specifications but that contain completely different items. Additional complexities arise because, to ensure equity, these forms need to be equated. Equating requirements can have a significant effect on test design. These considerations are beyond the scope of this chapter, but we refer the interested reader to Chapter 11, this volume as well as to Holland and Dorans (2006) and Kolen and Brennan (2004, Chapter 8).

Rather than go the expense of additional item development, an alternative is to use the same items but in a different order. Studies of item context effects (e.g., Kingston & Dorans, 1985) have shown that changing the ordering of items can change their difficulty. In their study of GRE test takers, these differences were usually small, and small increases in difficulty for some items were typically counterbalanced by decreases for others. This issue is still not well studied, so it might be important to equate forms for which the item ordering has been changed.

A third alternative would be to change the order of the options within items, regardless of whether

the items themselves were administered in the same order. There are reasons to believe that the order of options will make a difference in item performance (Haladyna, Downing, & Rodriguez, 2002), but this question has not been well researched. Impara and Foster (2006) provided a detailed description of item and test development strategies to minimize opportunities for cheating.

Design considerations when reporting group performance. When providing only group performance reports, there is no need to require that each test taker provide responses for the entire domain. A matrix sampling approach can be used, such as the balanced incomplete block design used by the National Assessment of Educational Progress and other programs (van der Linden, Veldkamp, & Carlson, 2004) to extend content coverage. Instead of testing each individual on the entire domain to be measured, there is the option of providing each test taker with a matrixed test that includes only a subset of the items to be measured for the entire group. These matrixed tests should be randomly assigned to test takers to preserve the validity of inferences regarding group performance. This approach will not reduce item development requirements, but it will decrease the testing time required for each individual test taker. An entire group, however, must complete the test before performance can be reported.

Design considerations when reporting both individual and group performance. Each test taker will need to be tested on the entire domain to be reported on individual reports. It is possible to include a matrix part of these assessments. Either the matrix part should cover the entire domain (but with increased item sampling), weighting should be used to maintain the domain specification, or the matrixed component should be used only to estimate group score distributions and not count toward individual scores.

This is not to say that the entire test must consist of items that cover the exact same content. If the content is sufficiently broad and deep that a single test taker could not be tested on items that cover every possibility, the matrix can be at a more specific level. For instance, a blueprint for a science test might specify that each form of the test have two items on the topic of wave properties of heat, light, and sound. Clearly, it would be difficult to have three subtopics within wave properties covered by only two items. However, there may be a need to report on how test takers do on all three subtopics. In this case, three forms of the test can be developed to cover the topic thoroughly: One form might test heat and light, a second form might cover heat and sound, and a third form might have questions on light and sound. Each test taker would then be tested on a random sample of subtopics from the superordinate topic, which is sometimes referred to as *random domain sampling.* An additional bonus is that for purposes of higher level reporting (such as a school district), there are now three times as many questions from that domain.

When using matrixed forms in this fashion, great care must be taken to ensure that the forms are parallel in terms of the topics covered (in the earlier case, each form had two items from the topic of wave properties of heat, light, and sound) and statistically equivalent at the level of reporting. It is highly unlikely that the three pairs of two items used in that earlier example will all have the same difficulty, but it is also unlikely that anyone would consider reporting on only two items. More likely, the test would provide an overall science score, in which case the test forms would only need to be equated at the total test score level. The test might also provide information on certain categories of science knowledge, such as physical science, life science, and earth science. In that case, the heat, light, and sound items mentioned earlier would be part of the physical science subtest, which could be equated to all other forms' physical science subtests.

Will the Test Be Used to Report Growth or Change in the Examinee's Performance Over Time?

The literature regarding the psychometric difficulty of accurately measuring individuals' growth is well established (see, e.g., Gulliksen, 1950; Lord & Novick, 1968). Accurately measuring the growth of groups is easier because error is divided by the square root of the sample size. Regardless, political pressures and parents' desire for measures of their children's academic

growth require attention, and different test development strategies can be used to approach this issue.

Statistical challenges to measuring change arise because in assessing change as the difference in performance on two occasions, true scores from the first occasion are subtracted, whereas error scores are added. Thus, the reliability (ratio of true score to true-plus-error scores) of a difference score is very low, regardless of the magnitude of the original measure's reliability. The only factor that ameliorates this issue is the correlation between what is tested on the first and second occasions. Perhaps counterintuitively, growth scores tend to be more reliable the lower the correlation between the constituent measures is.

The narrower the content measured, the more likely it is that the correlation between scores on the first and second occasions will be high and thus that the growth scores will have low reliability. The wider the content, the more likely the test will be to cover material that was not taught in class or addressed in therapy, meaning that the change is not closely related to educational or treatment programs of interest. A psychometrically better (but perhaps logistically or politically not viable) strategy is to increase the period of time between occasions, thus allowing a greater variability (and correspondingly lower correlation) between true scores on the two occasions.

Chapter 12 in this volume presents a much more complete description of the issues and approaches associated with measuring change. Regardless of the approach used to produce change scores, the reader should follow Lord and Novick's (1968, p. 76) exhortation to always provide an estimate of the variability (standard error) of an individual's change or growth score.

Who Will Get Reports? Which Reports Will They Receive? How Will Performance Data Be Categorized for Each Audience?

The test developer must not only determine all of the reports the test will support but also who will receive each report. This is important to make sure that the reports will be meaningful to the target audience. All of the information provided on the reports must be gathered, either with item responses from test takers or from survey information about the test takers. Make sure to gather all of the demographic data that will be needed for each audience to interpret scores.

Test results should be accompanied by an interpretive guide that helps examinees and other consumers of the data understand the meaning of the test results. A basic interpretive guide will dissect the report, section by section, and describe what each element means in plain language. A good interpretive guide will go a step further and provide information not only about what the results can tell readers, but also about what the results cannot tell readers as well as provide any needed caveats or cautions about overinterpretation of the test results.

Harvey (Volume 2, Chapter 3, this handbook) and Hambleton and Zenisky (Volume 3, Chapter 23, this handbook) provide more information about reporting results in psychological and educational settings, respectively.

There are additional questions the test developer should ask. How will the audience need to group performance? What subpopulations will be meaningful for this audience? If one is providing reports to people who will need to compare subpopulations, such as schools or ethnic groups, it is important to make sure that these data are available in the reporting system. If the data are not already available, then they will need to be gathered separately and merged with other data. If, for example, the reporting audience wants to group students by how many hours they spent preparing for the test, then that question would likely need to be asked in a survey, either at the beginning or at the end of the assessment. The challenges of merging data from different sources should not be taken lightly.

How Soon Will Reports Need to Be Made Available After Testing Is Completed?

The speed with which test results must be returned could have a big impact on the selection of the item types to include. This requirement may also have a big impact on the types of technology needed for administration and scoring. Finally, the answer to this question may lead to use of different statistical methods and procedures in developing tests.

If scores are needed right away, the blueprint should include only item types for which there is technology or an administrator to support immediate

scoring. Immediate reporting will not be possible if the blueprint includes item types that require responses that must be hand scored, unless such scoring can be done immediately by people available at the time of testing.

When simply reporting percentage or number correct for immediately scorable items, then scores can be provided right away for individuals. The provision of percentiles and percentile ranks will usually require delaying reporting until after normative data are first collected. If reporting equated scores, either a preequating method will need to be used (see, e.g., Kolen & Brennan, 2004) or it will be necessary to take the time to equate once data from the first administration of a new form are gathered. If reporting performance categories, standards setting will need to be conducted before results can be reported.

Tests for which immediate reporting is necessary require some additional up-front planning and extra quality control procedures. Generally, items are field tested before use; that is, items are administered to a group of examinees that is similar to the intended test-taking population. Field testing serves dual purposes: to weed out bad items that may be ambiguous, biased, or misaligned and to collect data on how the items perform psychometrically. See the subsequent section on field testing as well as Chapter 7, this volume. The data collected on the items during the field test can be used to create statistically equivalent forms, develop raw-to-scale score conversion tables using either classical or IRT scaling, and predict distributions of test scores for normative purposes. The amount of faith put into using field test results for these purposes should be a function of the severity of the stakes associated with the test score and the control one has over the field test administration.

If delivering the test via paper and pencil, rapid scoring and reporting may still be able to be provided if using a rubric and local hand-scoring option or, for selected responses, hand scoring, an optical reader (a scanner for bubble sheets), or a local scanning plus Internet option.

Should This Testing Program Be Administered on Computer or on Paper?

The choice of test administration approach—individual, group paper, group computer based—needs to be driven by program goals, not the other way around. The following considerations will help guide decisions.

Administration cost. On the surface, this consideration appears to be a no-brainer. Group administration of paper-based selected-response tests is very cost effective. Permanent administration centers are usually not needed. When all examinees can take the test at or about the same time, only a single test form might be necessary. Using this model, the PSAT is administered to about 3 million students in a single day at a cost (for the October 12, 2011, administration) of about $14 a student (as opposed to $160, the 2010–2011 price for the computer-administered GRE General Test, which is administered to 500,000 examinees a year).

However, other factors can lead computer-based testing to be a more cost-effective mechanism. Segall and Moreno (1999) reported on several cost analyses performed by the military to determine the economic feasibility of a computer-based Armed Services Vocational Aptitude Battery. It is important to note the wide variety of factors that affected cost, which included savings of administration time, the impact of flexibility in test starting times, and cost savings associated with selecting more qualified recruits (because of increased reliability).

Flexibility of administration dates and times. Individual administration and computer-based testing in dedicated centers provide the most flexibility for on-demand testing. Group-administered tests gain their cost efficiency by bundling as many examinees as possible on each test date.

Test reliability for a given administration time. It is often stated that computerized adaptive tests are more reliable than paper or computer-administered linear tests. This is a misstatement. One can always build a linear test of equal reliability to an adaptive test. However, computerized adaptive tests, whether the adaptivity is after each item or after a testlet (small set of items), are more reliable than a linear test per item or per unit time. Adaptive testing avoids administering noninformative items (items that are so easy or so hard that they add little to

what one knows about the examinee). A cost analysis is required to determine the value of the time savings for a specific testing program.

Item types. Computers can administer and score item types that are hard or expensive to administer and score on a paper-based test. Use of non–selected-response item types can be desirable for many reasons. In some cases, the validity arguments for the construct of interest have been seen to require free-response items, as has increasingly been the case for measures of writing proficiency. The abilities to recognize good writing and produce good writing, although related, are not the same.

Another common validity relates to the back-solving issue for certain quantitative problems. Consider the problem of solving a quadratic equation. A quadratic equation takes the form $ax^2 + bx + c = 0$ and is important in several areas of applied mathematics. There are two primary ways to solve quadratic equations—factoring and using the quadratic formula, which is

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Consider a multiple-choice item asking for the solution of the following relatively simple quadratic equation:

> For what values of x is the following quadratic equation correct?
> $4x^2 + 2x = 0$
> a. $-\frac{1}{2}$ and 0
> b. $-\frac{1}{4}$ and 1
> c. 0 and 1
> d. 0 and 2

Rather than solving the quadratic equation in a way in which one would need to if there were an infinite number of possible answers (as is the case with many mathematics problems), there are only five possible values that need to be checked for correctness. Plugging each of the five values into the left-hand side of the equation quickly allows examinees to determine the correct answer regardless of their ability to factor or to know and be able to use the quadratic formula. Back-solving quadratic equations is a different construct than solving them.

Other times, because teachers often model instruction after items in high-stakes tests, consideration of the instructional consequences of using only selected-response items may lead to inclusion of more complex item types.

In some cases, the (at the time) greater efficiency of using scannable documents moved some personality and interest assessments away from their original sorting method (see Block, 1961/1978) to a two-or-more-choices selected-response method. Computers could readily support such sorting methods.

Regardless of the reason for using non–selected-response item types, their inclusion will require several decisions. Some testing programs choose to use human graders. To ensure consistency, this will require careful preparation of scoring guides (rubrics) and often exemplar responses and training materials. Sometimes human graders are available at no expense to the program (such as the historical New York State Regents Examination program, in which classroom teachers were required to grade their own students). Other times, grading sessions are looked at as a professional development activity and held at a central location with value that compensates for the expense.

An alternative to human grading is computerized scoring. Depending on the type of question, different approaches have been used (Koul, Clariana, & Salehi, 2005; Williamson, Almond, Mislevy, & Levy, 2006).

Items that require a numerical response can be scored by matching against the correct answer. Acceptable ranges of responses can be used if the problem requires them (e.g., any answer between 3,996 and 4,004 is counted as correct). The scoring engine can be set to reduce answers into a common form so that $3\frac{1}{3}, \frac{10}{3}$, or any number between 3.33 and 3.3333333 is counted as correct.

Responses to short-answer questions can be checked against a list of acceptable responses, which might include common misspellings. Drag-and-drop questions (a computer-based item type for which users must drag objects to the appropriate part of the screen, such as placing astronomical objects into their appropriate place in a sky map) can be matched against the boundaries of acceptable placement. Computerized sorting of items would be one form of drag-and-drop item.

One common method used for the computer-based scoring of essays is latent semantic analysis, which is a statistical approach that, in essence, compares the use of words in the essay being scored with those in a target packet of previously scored essays and predicts what score would be given by a human grader. One concern with latent semantic analysis is whether, with a little training, examinees could be coached on how to improve their score without a concomitant improvement in the quality of their writing (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). This problem is unlikely for low-stakes tests because the incentive for such cheating is low.

Should Examinees Have Limited Time to Take the Test?

Time can be a factor in test administration for two reasons. Often the logistics of test administration—issues such as administration center availability and cost—can limit the time available for a testing session. Alternatively, the construct of interest may be inherently time dependent (such as typing speed).

The amount of time available for testing will have a significant impact on the number of items able to be presented to each test taker. Reading time—or, more generically, time to interact with the stimulus—must be included in the estimated testing time. More than one test has run over the predicted testing time when the time to read passages, scan a map, or read a table was not considered in the estimates. If one wants to include the time that test takers require to answer questions as part of the construct, the testing time should be tailored so that the time-proficient test takers will complete the assessment and the non–time-proficient test takers will not. One must keep in mind that average reading speed varies with age. Although initial estimates for each item should be made during development (on the basis of the number of words and complexity of the tasks), it is usually a good idea to time test takers during the pilot- and field-testing stages of development to get a more accurate estimate of testing times.

One rule of thumb for speededness is based on (a) the last item to which 100% of the examinees responded (expressed as the percentage of items) and (b) the percentage of examinees who responded to the last item. For many decades, the Educational Testing Service used as a rule of thumb that a test was speeded if 100% of the examinees responded to fewer than 75% of the items or fewer than 80% of examinees responded to the last item (Swineford, 1974). These measures assume examinees move through a test in a linear fashion, which is increasingly unlikely to be true given the proliferation of test preparation books and courses that advise against that test-taking strategy.

Even if the test will not be timed, the testing time should be estimated. Test users will want to know how much time to allocate to testing for logistical planning. However, it is often not a good idea to have a completely untimed test because some test takers will perseverate well beyond what is necessary.[2]

If building an adaptive test, testing time may vary dramatically across test takers. This variation may be caused by significant differences between initial ability estimates and actual test-taker abilities. Test-taker times may vary for diagnostic testing by as much as 5 hours when the student's assumed level was significantly above or below the student's actual ability. The trade-off in such cases is reliability of the report versus the number of items or tasks that must be presented to the test taker. If a significant variation in test-taker ability is expected, the test may have the same issue.

What Are the Stakes? How Likely Is Cheating?

If high-stakes decisions will be made on the basis of the test, the test will probably need to be administered in a secure, proctored environment. Proper proctoring is not intuitively obvious. The test should have a manual that describes appropriate security for all stages of the testing process, including shipping, storage, administration, and retrieval or destruction of secure materials after the administration.

[2]When Neal M. Kingston was associate commissioner for curriculum and assessment for the Kentucky Department of Education, he was made aware of a student who spent 4 days responding to constructed response questions expected to be answered in 1 day. The student's answers were excellent but went well beyond what was required to demonstrate excellence.

If the test is to be a secure test, one must make sure that the items are also secure during development. For example, in a state testing program, delivering the test in a secure environment is of little use if the lack of security during the development process led to many teachers having copies of the tests in their drawers.

If the stakes are high enough that cheating is likely to become a significant issue, multiple parallel forms can be created of any linear tests being developed to help address this issue. It is also a good idea to have a form in reserve in case one or more forms of the test are exposed (either intentionally or accidentally) before administration. Providing multiple forms with items in differing orders can help minimize cheating if test takers will be sitting next to each other during administration. When doing this, developers must make sure not to cause item ordering differences between test takers on item sets. For example, reading passages often take advantage of natural scaffolding: The more global items (such as main idea) are generally presented first, with more detailed questions coming later in the set. It may be best to reorder the sets, rather than the items within a set, if item order is an issue for the content.

If using a computer adaptive test, item exposure (how often each item has been administered) should be monitored as well as any changes in item difficulty, because item exposure increases over time. Item pools should periodically be replaced or refreshed.

How Transparent Should the Testing Program Be?

Professional guidelines such as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) require that information be made available to various constituencies involved in the testing process. At the very least, a test publisher needs to consider how and when to produce administration manuals, explanations of score reports, and technical documentation.

Higher levels of transparency, including periodic release of intact test forms, are sometimes required by law (New York Education Law, 1979–1980). Even when not required by law, high levels of transparency might be desirable to convince the public of the appropriateness of a testing program.

Releasing test forms, whether periodically or regularly, requires planning. In the extreme, if every form of a test will be released, item data sufficient to support preequating must be gathered before those items are publicly exposed. This can be done by embedding operational test items that do not count toward student scores.

An alternative approach that has often been discussed but, to our knowledge, has not been implemented, is to release a pool of items so large that having access to the items does not increase student scores to any appreciable extent. A few research studies have looked at this approach with inconclusive results (Hale, Angelis, & Thibodeau, 1983; Powers, 2005; Powers & Fowles, 1998; Powers, Fowles, & Farnum, 1993).

DEVELOPING THE TEST

After answering the appropriate strategic questions and developing a test blueprint, the heart of the work begins. This section covers item development, item editing, content and access review, bias review, sensitivity review, pilot testing, field testing, form and item bank creation, standard setting, and creating an administration manual.

Item Development

Select item writers who have content-specific knowledge. Although there are tips and techniques to writing very technically sound questions that will have strong psychometric properties, knowing the right questions to ask is of greater importance. People with content knowledge can usually be trained in good item-writing practices. If not, item editors can take care of these issues (although a final review by content experts is usually necessary to ensure that meaning has not unknowingly been changed).

Item writers should be provided with detailed item specifications. These specifications will include those for stimuli (e.g., reading passages, tables, graphs, video, audio, images) as well as the item types and rubrics, as needed. If using distractor

analysis in reports, specifications should also include information as to how to define the meaning of the incorrect answers. These specifications will be used to select or write new content.

If a large number of items are to be written for specification, then it may be advantageous to develop item templates that can be used for either fully automated or semiautomated item development. Item templates should include constraints on the stimulus, stem, and answer choices or rubric, depending on the type of item to be developed.

It may also be useful to include learning maps, learning progressions, or scope-and-sequence information on these templates, which will enable item writers to select meaningful distractors that will differentiate prerequisite from target knowledge for the purposes of distractor analysis in the reports. Distractors may also be selected that differentiate common errors or misconceptions from correct responses in the target knowledge or skills that are unrelated to the prerequisites. In some cases, it may also be beneficial to include distractors that identify interference from more advanced skills.

At this stage, item writers must be aware of universal design principles and any accommodations that will be supported. Although universal design will improve the item's accessibility for various populations, it will not completely remove the need for some accommodations. The use of accommodations should not interfere with the construct being measured, so the item writer should know what accommodations will be allowed. Whenever possible, item writers (and test developers) should, while taking advantage of technological solutions or one-on-one testing environments, include additional information on, for example, how an item could be read aloud without giving away the answer. In some instances, such as when translating an item into American Sign Language, external experts may be required to work with the test developer to ensure that the item can be presented and answered in a way that supports valid interpretations of examinee performance.

Item Editing

Once the content is written, it is edited for accuracy, style, universal design, and anticipated bias and sensitivity issues. Ensuring content validity at this stage is made easier by comparing the items with the item specifications in the blueprint, which allows the editor to easily determine whether the desired measurement was achieved from a face validity perspective. See Chapter 16 in this volume for more detail on item editing.

Content and Access Review

Each item should be reviewed by more than one individual to ensure construct clarity, accuracy, age appropriateness, and consistency of style. Content and access review panels should consist of subject matter experts as well as representatives from historically disadvantaged groups. Although a subject matter expert will know the right questions to ask, someone who works with individuals with disabilities or English language learners will be able to point out issues with the items that might interfere with the examinee's ability to respond to the test question. The two groups—content experts and specialists—should work together to improve accessibility without compromising the integrity of the content.

Bias Review

Items should be reviewed by more than one person to ensure that the test as a whole will be appropriately balanced for cultural, socioeconomic, and gender perspectives. Bias review panels should consist of representatives from subpopulations of interest, such as ethnic groups; those from an urban or rural setting, geographic region, or high- and low-poverty areas; teachers of students with various special needs; and other groups that might be historically or potentially disadvantaged or underrepresented.

Teresi and Jones (Chapter 8, this volume) present details regarding fairness issues in psychological testing, and Zieky (Chapter 17, this volume) presents details on procedures used to minimize the impact of item bias.

Sensitivity Review

Although many test developers agonize over content and construct validity, too little concern is often given to public (and political) sensitivity issues. Sometimes a test must be controversial. For example, the assessment of student knowledge about

evolution has been controversial in several states, but if evolution is part of the life sciences test specification (and it should be) it must be assessed. However, if a controversial issue is not part of a test specification, using that issue as a context for test items is at best a distraction to students and at worst a lightning rod for attacks on education. An example of this latter problem might occur in a state in which there is conflict between the lumber industry and environmental activists, and a math problem is set in the context of determining how many pages of a certain size are needed to publish an environmental newsletter. Similar sensitivity issues can be associated with psychological tests and interest inventories.

It is also important to recognize that certain forms of testing require items that many constituencies will find broach sensitive topics, for example, tests of racism (see Volume 2, Chapter 25, this handbook).

Sensitivity issues are often issues of balance. Whether a contextualized test item has a male or female character is usually not a problem. However, it might be a sensitivity issue if every item is about a girl and none are about boys or if all characters are chosen to be counter to traditional stereotypes (e.g., all women are business executives and all men are homemakers). If most of the items about U.S. history focus on the history of only part of the country, it might be a sensitivity problem. Sensitivity issues differ for individual testing programs but can be minimized or defended against by virtue of the review process used.

Pilot Testing

No fixed agreement exists in the measurement community regarding the terms *pilot testing* and *field testing*. We use *pilot testing* to describe a small-scale activity and *field testing* to describe a larger activity likely to yield statistical results that have a high degree of precision.

Whenever possible (determined by both cost and availability of appropriate samples), items should be given to small groups of test takers to determine whether any interpretation differences exist between the developers of the items and the test takers. Observing how test takers interact with items during this stage can indicate whether directions are clearly stated, whether editing changes have made the items too wordy, and whether editing has left items ambiguous or missing necessary information as well as the amount of time needed for a test taker to respond to an item. Piloting items is especially important if a new item type or delivery technology is being used. Any issues detected here should be corrected before field testing. If significant changes are made, it would be good to revisit the item in a content and bias review process.

Field Testing

Experience has taught us that even the most experienced, most skilled test developers make mistakes. Items that seem completely unambiguous to an expert might not be so clear to novice learners. Also, even with a rigorous review process, the sheer number of items developed for a large testing program makes it inevitable that some flawed items will make it through review processes based solely on human judgment. Moreover, even good items might be more or less difficult than content experts thought, and thus the test might not measure as accurately as desired at key parts of the score scale. For these reasons, items should be field tested with a representative population.

There are two main types of field testing—stand-alone and embedded. A stand-alone field test is administered separately from the operational or "real" test, and embedding is done by inserting field test items into a real test that is currently being administered. Each approach has its advantages and drawbacks. Embedded field testing, in which examinees do not know which items count and which do not, has the huge advantage of equal student motivation on the try-out items. Thus, item statistics for new questions are comparable to what would be expected in their operational usage. Embedded field testing also makes it easier to attain a representative sample because the field-test items can be randomly distributed to all test takers. Stand-alone field testing, if not done on a census basis (testing everyone), relies on purposively drawing a random sample of test takers. In either case, a random test form distribution method is required to ensure that groups of field-test items are randomly equivalent.

Embedding is also logistically simpler because only one testing event is needed. When testing is expensive, takes time away from job duties or classroom activities, or is politically charged, the goal with testing is to get in, get the data, and get out as quickly and unobtrusively as possible. With stand-alone field testing, data collection is a separate event that, although it may be on the same day, more commonly occurs before or after the operational test, sometimes offset by weeks or months.

This time differential can be very important if what is being measured is expected to change over time or with treatment, training, or other intervention. For instance, student performance on a test administered at the end of a U.S. history course would be expected to exceed student performance on that same test if it were administered halfway through the course. If the stand-alone field test is conducted just after the teacher has finished teaching about World War I, test-taker performance on items dealing with World War II will be substantially lower. If those item data are used to create pre-equated forms, establish score scales, or set cut scores, it will create an inaccurate picture of student performance when the tests are administered operationally at the end of instruction. Items that seemed to be very difficult when the material was not yet taught will suddenly be much easier when the material has been taught.

Although the advantages of embedded field testing are very obvious, under certain circumstances, stand-alone field testing is the better choice. Stand-alone field testing is often done in the 1st year of a new testing program because there may not be an operational test into which field-test items can be embedded. If an existing testing program is changing, there may be significant enough changes to item content, item or test format, or test delivery to make an embedded model less useful. If a test taker can identify which items are field-test items and which are live items, the main advantage of using embedding is moot.

If the quality of embedded field-test items is questionable, it is possible that a flawed item could disconcert examinees and throw off their performance on other items. Initial pilot testing can minimize this, and post hoc analysis can show the extent to which such concerns appear valid.

Stand-alone field testing has another advantage in that it can be used as a model for how an entire full-length test will behave. Embedding is sometimes done with a few field-test items sprinkled here or there, and sometimes an entire section of the test is composed solely of field-test items. Although embedding can be continued while making subtle changes to the items, knowing to what extent the changes will compound when delivering a test composed only of the new items is not always possible. For instance, say that one wants to increase the cognitive demand of items in a test, moving away from simple recall and more toward integration of ideas at a higher level. If five or 10 of these items are field tested in a current operational form, one may or may not notice that test takers need slightly longer to complete the test. Once an entire form is made of these new items, one may find that test takers now need significantly more time than what was originally set for the testing time limit.

At this stage, delivering the items in forms with the same content balance as that of the final form is not necessary. In fact, it may be beneficial to field test more items for content areas or item types that are more likely not to meet psychometric requirements.

Form and Item Bank Creation

Using the statistical information from the field test, select the best items for the forms (see Chapters 7 and 10, this volume, for more information on this topic). When creating multiple parallel forms from the items in the field test, it is important to select the forms simultaneously. This is necessary to avoid building the first form using all of the "best" items, because the reliability of each of the following forms would suffer. It is better to distribute the best items evenly across the forms so that they will all have about the same reliability. Depending on the method used for reporting scores, statistical information such as *p* value (the percentage of students who answered an item correctly), the point-biserial correlation (the correlation between answering the item correctly and total score), the point-biserial correlation for each of the response options (incorrect responses should have a negative or low

point-biserial correlation[3]), and perhaps IRT parameters (*a, b, c*) will be used. There may also be statistics for differential item functioning that provide the information for a statistical bias review. If an item has high differential item functioning, it is more likely to be answered correctly by test takers within a given subpopulation of interest. In addition, there may be correlations between the item and subscores as a statistical basis for content review.

When building a CRT, the test forms developed must meet the test blueprint in terms of content to be assessed. When building a math test that requires assessment of numeracy, algebra, and geometry, there should be items measuring all those traits. It is possible to get "bad" statistics on those items, particularly if a topic is new to the content area. However, to make interpretations of the data relative to the content, as done in a CRT, it is crucial to ensure that all the content is being assessed. When seeing a cluster of items with bad statistics, one must look at the items. A pattern might be found in those items that can inform either subsequent item development or needed changes in the preparation of test takers. In the case of a CRT, one must be sure to communicate to the relevant stakeholders that test takers may require additional instruction or training on certain content. For example, a state's math test had items requiring students to determine the perimeter and area of plane figures. Half of the items field tested had really good statistics, and half had really awful statistics. On examination of the items, it quickly became apparent that the students were doing fine on determining area and perimeter of regular and convex polygons, but that items that had irregular or complex polygons were giving the students fits. Simply writing more items in the hopes that some of them would have better statistics would be a very weak solution to this problem, and including only items that use regular and convex polygons would be neglecting part of the content. A quick conversation between the state's test developer and the state's math instructional leadership resulted in professional development for the state's teachers.

Standard Setting

Standard setting, although a critical part of the development of many tests, is not described in this chapter. Volume 3, Chapter 22, this handbook, is dedicated to that topic.

Test Administration Manual

Once the test has been developed and the administration methods determined, an administration manual must be developed. A test administration manual is important to ensure standardized test delivery to all examinees. If one group of examinees has a time limit for taking the test and another group of examinees does not, the interpretations drawn from the two groups of data will not be consistent. Generally, such a manual is developed as items are piloted to ensure clarity and consistency in the instructions to test takers and test administrators. When field testing items, the manual can also be field tested.

This manual should provide background on the assessment, instructions to test administrators and proctors (if any) about the procedures to follow when delivering the test (e.g., delivery order, timing, or technology requirements), and guidelines for providing accommodations and methods for scoring the test. In addition, it should explain how to access and interpret the reports unless a separate interpretive guide is also developed.

PLANNING TESTS OVER TIME

In test development, it is often very beneficial to think of the entire life of the assessment over time. Many times, a given test cannot be released with all of the features one would like to have in the first administration, and even if it could, one may want to refresh the test over time for security reasons or because of changes in the standards. Planning is key.

Time is an assessment program's worst enemy, but it can also be its best friend. Development takes time, and features in a program will often be cut because of the short calendar time between specification and the operational date. With careful planning, however, the assessment can become more effective over subsequent releases.

[3]Some psychometricians (including Neal M. Kingston) recommend against using the point-biserial correlation for incorrect options because the correlation coefficient assumes a linear relationship that should not expected for any but the correct option. Nonetheless, this approach is commonly used.

Some who want to do adaptive testing do not have the time, money, or available test takers to field test enough items for an adaptive test in the first release. By embedding items in each administration, an item pool can still be built up for use in future administrations. Online testing could provide the most flexible field-testing options. If the technology is available, embedded matrix field testing may be possible. Field testing in an embedded matrix design allows testing of more than a single set of items. Items can be included in field-test slots until enough of the population has taken them to support the required statistical analyses, at which point other items can be put into those slots. Because field testing need only be against a representative sample of test takers, many more items can usually be field tested this way.

Other changes that can take place over time within a testing program include the addition of new forms. As mentioned earlier, having several parallel and equivalent forms to rotate through, as well as an extra form, may be beneficial in the event that the security of one of the live forms becomes compromised. For example, what if initial field testing did not have enough items survive to make as many test forms as wanted? What if there is a catastrophic security breach and multiple forms are lost? What if the same form or forms of the test have been given so many times that test administrators can recite the test in their sleep? In those cases, the supply of forms will need to be replenished.

Yet another issue is how to ensure that new forms are equivalent to the old forms, particularly in an assessment environment in which change is expected to occur. This is commonly called *equating,* and several methodologies exist to accomplish this (see Chapter 11, this volume).

DOCUMENTATION

Earlier, the notion of face validity was presented as well as the idea that much mistrust of assessment programs comes from the layperson's lack of understanding of what goes into sound test development. Each step of the test development process needs to be documented and made available to researchers, stakeholders, the press, and other interested parties. At a minimum, good test documentation, such as a technical manual, should include information on the purposes of the test, how the content to be tested was decided on, the test development process (and the many layers of review), information gleaned from the pilot and field testing, the results of the live administration (such as mean scores, standard deviations, and reliability), the standard-setting method and results, and a discussion of sources of validity evidence for interpretations of the test score. The documentation should be supplemented and updated after each live administration of the test forms. Chapter 14 in this volume provides more information on this important topic.

CONCLUSION

To develop the best possible test, start by defining the inferences test takers should be able to make. Next, determine the strategies that will best support those inferences. Finally, using this and other chapters in this handbook, choose and implement the tactics used to carry out the plan.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69,* 49–56. doi:10.1111/j.2044-8325.1996.tb00599.x

Block, J. (1978): *The Q-sort method in personality assessment and psychiatric research.* Palo Alto, CA: Consulting Psychologists Press. (Original work published 1961). Retrieved from http://www.qmethod.org/articles/jack_block.pdf

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511571312

Dana, R. H., Bolton, B., & Gritzmacher, S. (1983). 16PF source traits associated with DSM-III symptoms for four diagnostic groups. *Journal of Clinical Psychology, 39,* 958–960. doi:10.1002/1097-4679(198311)39:6<958::AID-JCLP2270390622>3.0.CO;2-7

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440. doi:10.1146/annurev.ps.41.020190.002221

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley. doi:10.1037/13240-000

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309–333. doi:10.1207/S15324818AME1503_5

Hale, G. A., Angelis, P. J., & Thibodeau, L. A. (1983). Effects of test disclosure on performance on the Test of English as a Foreign Language. *Language Learning, 33*, 449–464. doi:10.1111/j.1467-1770.1983.tb00944.x

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Impara, J. C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Mahwah, NJ: Erlbaum.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Author.

Kingston, N., & Dorans, N. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*, 281–288. doi:10.1177/014662168500900306

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Koul, R., Clariana, R. B., & Salehi, R. (2005). Comparing several human and computer-based methods for scoring concept maps and essays. *Journal of Educational Computing Research, 32,* 227–239. doi:10.2190/5X9Y-0ETN-213U-8FV7

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

New York Education § 340 et seq. Standardized testing (McKinney Supp. 1979–1980)

Powers, D. E. (2005). *Effects of preexamination disclosure of essay prompts for the GRE analytical writing assessment* (GRE Board Report No. 01-07R). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping *e-rater*: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*, 103–134. doi:10.1016/S0747-5632(01)00052-8

Powers, D. E., & Fowles, M. E. (1998). *Test takers' judgments about GRE writing test prompts*. (GRE Board Report No. 94-13R). Princeton, NJ: Educational Testing Service.

Powers, D. E., Fowles, M. E., & Farnum, M. (1993). Prepublishing the topics for a test of writing skills: A small-scale simulation. *Applied Measurement in Education, 6,* 119–135. doi:10.1207/s15324818ame0602_2

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Schneider, L. M., & Briel, J. B. (1990). *Validity of the GRE: 1988–1989 summary report*. Princeton, NJ: Educational Testing Service.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 35–65). Mahwah, NJ: Lawrence Erlbaum Associates.

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26,* 21–28. doi:10.1111/j.1745-3992.2007.00105.x

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81–97. doi:10.1037/1082-989X.1.1.81

Swineford, F. (1974). *Test analysis manual* (Statistical Report SR-74–06). Princeton, NJ: Educational Testing Service.

van der Linden, W., Veldkamp, B. P., & Carlson, J. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28,* 317–331. doi:10.1177/0146621604264870

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case of testlets. *Journal of Educational Measurement, 24*, 185–201. doi:10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., . . . Thissen, D. (2001). Augmented scores—"Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–388). Mahwah, NJ: Erlbaum.

Williamson, D. M., Almond, R. G., Mislevy, R. J., & Levy, R. (2006). An application of Bayesian networks in automated scoring of computerized simulation tasks. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 201–257). Mahwah, NJ: Erlbaum.

CHAPTER 10

ITEM BANKING, TEST DEVELOPMENT, AND TEST DELIVERY

David J. Weiss

The paper-and-pencil (P&P) test that dominated psychological, educational, personnel, and other applications of testing for the majority of the 20th century was born in the second decade of the 1900s in response to the personnel needs of World War I (Dubois, 1970). With the need to screen and classify large numbers of recruits rapidly and efficiently, the then-predominant mode of testing by individual psychologists was not able to meet the demands of the U.S. military. The multiple-choice test question was invented, and tests were written and printed and given to groups of recruits—the first major implementation of group, rather than individual, testing.

Because of its efficiency, P&P testing spread rapidly into other fields that had previously relied on individually administered tests—education, intelligence testing, and other personnel testing applications. P&P testing also began to be used for measuring attitudes, interests, and other personality variables, thus permitting the recently born field of psychology to generate data on a wide variety of variables quickly and efficiently.

Although data acquisition using P&P tests was efficient, the process of test development—especially for larger testing programs—was anything but efficient. Figure 10.1 provides an overview of the major components of the test development process. For at least the first 50 or 60 years of P&P testing, maintaining a collection of items for any continuing testing program, including classroom testing, was a tedious process fraught with numerous opportunities for error. Test questions (items) were frequently written, or perhaps typed, on index cards. The cards were kept in file drawers, sometimes separated into content classifications. When item statistics were available for items, they were frequently written on the backs of the cards, identified by test form and date. To create a test, the test developer would manually search the file drawer, review the content and statistics for an item, and put it aside if selected for use in the test. When a sufficient number of cards had been selected, they might be reviewed by others, and some replaced with another card from the file drawer while the rejected items were returned. There were obviously many opportunities for item cards to get lost, misplaced, or misfiled.

Once an appropriate set of items had been selected, the cards would be manually put in the desired order and then typed onto a duplicating master. If an alternate form was needed, the order of the cards would be modified and a new test would again be typed from the cards. Of course, the typed test forms had to be proofread each time to ensure that the text of the test items had not inadvertently been changed from that on the index cards. The next steps, which are still necessary today for P&P tests, were duplication, collation, shipping (if required), distribution to the examinees, collection of answer sheets and booklets after administration, and scoring. Between the mid-1930s and through the end of World War II, only the largest testing programs had access to machines that could scan

DOI: 10.1037/14047-010
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

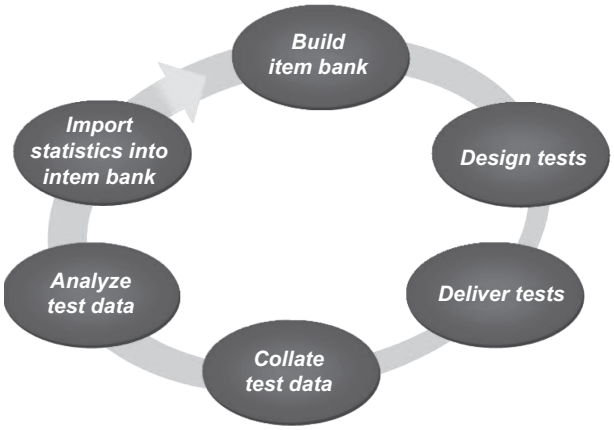

FIGURE 10.1. Major components of the test development cycle. Copyright © 2010 Assessment Systems Corporation. Reproduced with permission.

the answer sheets and provide scores. The alternative for the vast majority of testing programs was hand scoring of the answer sheets: A template was placed over each answer sheet and the number of marked answers was manually counted—a procedure also fraught with the potential for error. Moreover, the entire paper-based process was obviously inefficient and time intensive.

IMPACT OF TECHNOLOGY ON TESTING

As with many aspects of people's lives after World War II, technology began to have an impact on testing. The first impact was on the error-prone and labor-intensive test-scoring process. Less expensive optical scanners began to appear in the early 1950s that were capable of reading answer sheets, comparing the scanned answers with a set of correct or keyed answers and producing a score (and in some cases multiple scores) for each answer sheet scanned. The early machines were very large and expensive devices that were not true computers but provided some basic functions that were similar to those of computers—they were, effectively, "business machines" designed for a specific purpose. They were slow and temperamental and sometimes had reliability problems, but they were considerably more efficient, less expensive, and likely more accurate than the hand-scoring process using templates. Because of their expense and temperament, the machines were maintained by specialized staff, and answer sheets had to be mailed to the scoring organization and results mailed back to the test user. Thus, this process eliminated the labor necessary to hand score answer sheets but created delays both in the transport of the answer sheets and in their processing at the busy scanning centers.

Initially, these machines did not provide any group summaries of results. If an item analysis was desired, the item responses could be output on punch cards, and the cards could be run through other business machines to obtain basic frequency counts that could be used to hand-compute classical item difficulties. Other statistics, such as point-biserial correlations, would have to be computed by hand using a calculator. Of course, the test development cycle would have to be completed by hand-entering the item statistics onto the backs of the index cards for review by the test developers so that poorly functioning items could be identified for exclusion or revision before use in future tests.

Computers began to be used in testing as they became more generally available in the early 1960s. Their first application was to replace the early scanners, providing somewhat more reliable scanning, faster scanning, and the capability to be programmed to incorporate item analysis results for a defined set of answer sheets. Computers also began to be used in that decade in some larger organizations (particularly universities) for more complete test analysis, including validity analyses and factor analysis. Although the early vacuum tube computers were somewhat unreliable and used rudimentary input–output devices (e.g., punched paper tape output), the introduction of the solid-state computer and more reliable input–output equipment improved their performance considerably. For the first time, psychometric analysis could be done without hand calculations, but the rest of the test development cycle still remained the same in 1970 as it had been for the past 50 years since the inception of the P&P test.

Minicomputers came on the scene in the mid-1970s as solid-state computers began to shrink in size. These computers were one-tenth or less the size of the original solid-state computers of the previous decade and extended computing power to many organizations and projects that did not have it

previously. Their impact on testing was relatively minimal, with one exception noted later, except for making scanning and basic item analysis more widely available and less expensive.

PERSONAL COMPUTERS' IMPACT ON TESTING

Major changes in the way tests were developed, analyzed, and delivered began to occur with the introduction of the personal computer (PC) in the mid-1980s. This impact can be divided into three phases—storing items, banking items and assembling tests, and delivering tests.

Storing Test Items

As a labor-saving device for the production of manuscripts and other documents, the PC came with word-processing software that could be adapted for other purposes. Thus, one of the first uses of the PC in testing was to allow test developers to store test items in word-processing files. Word-processing software allowed test developers to type their items only once, then select them as needed to create a test. A new document could be opened and the text of stored items copied and pasted into the new document in the order in which the items were desired in the test. This process effectively removed the necessity to completely proofread a new test assembled from the master test-item files and made it easier to assemble alternate forms of tests when needed. All that was required of the test developer after a test was assembled was to check to see that all of the items were in the correct place and that items did not break across pages when the test was formatted—a considerable time savings from proofreading and correcting one or more forms of a test.

Some test developers adapted other standard PC software—notably spreadsheets—as item storage mechanisms. Again, the advantage was that item text could be stored, copied, and pasted to eliminate retyping. An added advantage was that the different pages of the spreadsheet could be used to separate items into subsets, perhaps representing the structure of a domain to be tested. A final advantage was that different cells in the spreadsheet could be used to store other data on the items and that information could be physically associated with the items, thus allowing both sides of an item "index card" to be stored together. Although word processors also allowed storage of the full index card for an item, spreadsheets were more flexible in their layout options and could more easily hold a wider variety of information on an item. Both word processors and spreadsheets allowed users to search for items that had specific values of an item statistic but were generally limited to simple searches.

Item Banking and Test Assembly

The test development process improved dramatically as special-purpose software was developed for item banking and test assembly (Vale, 2006). In the development of this type of software, an item bank was conceptualized as a database, and database software was programmed to perform the special functions necessary to maintain a testing program. Item text became one or more fields in a database, additional fields were defined for item statistics and other information, and the development of hierarchical structures to represent bank structures was facilitated by the item-banking software.

DOS item-banking software. Specialized item-banking software using the DOS operating system on PCs first began to appear in the mid-1980s. These text-based bankers generally had very limited graphics capability because computer displays and printers of that era were primarily text oriented. Thus, with the exception of the MicroCAT Testing System (Assessment Systems Corporation, 1987), which had relatively advanced graphics for its era, item bankers in the late 1980s were limited to tests that used items consisting almost entirely of text. They also had very limited formatting capabilities, in terms of fonts and special effects, because these were not supported by the line printers available at that time.

Nevertheless, the DOS-based item bankers greatly improved the efficiency of test development. Some had search capabilities that allowed the test developer to search through stored item information to identify items that had specified sets of characteristics—frequently permitting searching on multiple variables simultaneously—to identify candidate items for a specific test. Some permitted limited test formatting

and page numbering, and many permitted the easy creation of multiple forms of a test.

At the same time that item bankers were simplifying and streamlining the storage of items and item statistics and permitting test developers to assemble tests with specific characteristics, some special-purpose software extended test-assembly capabilities even further. Most notable among these was ConTEST (Timminga, van der Linden, & Schweizer, 1996), which was designed to solve complex test-assembly problems that involved a large number of constraints. ConTEST used linear programming methods to create one or more tests that satisfied all the constraints imposed. It operated from item statistics, however, and the resulting tests had to be manually assembled from separate bankers or databases.

Windows item bankers. Although the DOS-based bankers began to change the way items were stored and tests were assembled, they were quite rudimentary compared with the Windows item bankers of the 21st century (e.g., PARTEST [http://www.scantron.com/parsystem/], LXRTEST [http://www.lxrtest.com/site/home.aspx], and FastTEST [Assessment Systems Corporation, 2010b]). Windows item bankers usually incorporate a complete point-and-click interface to allow the test developer to interact with a database structure designed specifically for purposes of item banking and test assembly and can incorporate a range of types of graphic displays in items. The ability to print tests with these bankers was greatly enhanced by the widespread availability of PC-compatible laser printers, beginning around 1990.

The most useful item bankers allow item banks to be designed to reflect the structure of the domain to be tested, which is frequently operationalized in a test "blueprint." The blueprint is usually an outline or a hierarchical structure that delineates the structure and subdomains of the primary domain, frequently with additional levels of specificity. The number of levels in a bank hierarchy is determined by the structure of the domain, sometimes combined with characteristics of the test items. Figure 10.2 shows an item bank structure from the FastTEST Test Development System (Assessment Systems Corporation, 2010b) for an introduction to psychological measurement item bank.

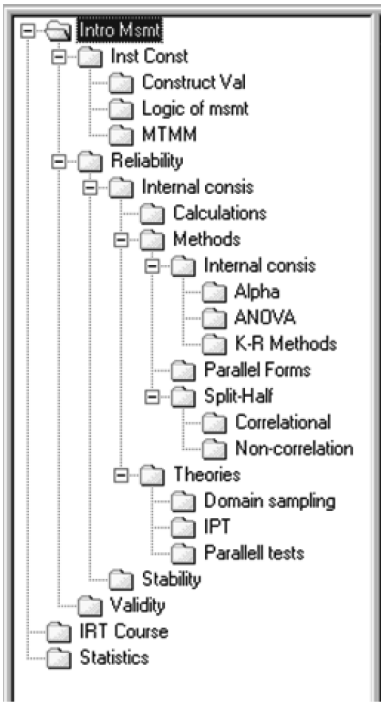

FIGURE 10.2. Bank structure of an introduction to psychological measurement item bank. Copyright © 2006 Assessment Systems Corporation. Reproduced with permission.

In addition to storing item text and any related graphics, item bankers allow storage of other information associated with each item. This information will, of course, include item statistics. Some bankers are designed only for use with classical test theory statistics—item difficulty (proportion correct) and item discrimination (biserial or point-biserial correlation)—and others allow storage of item parameters from item response theory (IRT; see Chapter 6, this volume) and display of IRT item functions (e.g., Figure 10.3).

Other information stored on items includes the correct or keyed answer to the item, tests in which it has been used, name of the item writer and date created, special user-supplied statistics (e.g., Angoff rating), keywords that characterize the items, and other notes concerning the item. The information associated with each item is typically organized in a set of tabs for easy access. For example, FastTEST has five tabs: Item Identifier (including keywords and description), Item Text (using a full-featured built-in word processor), Item Information (item

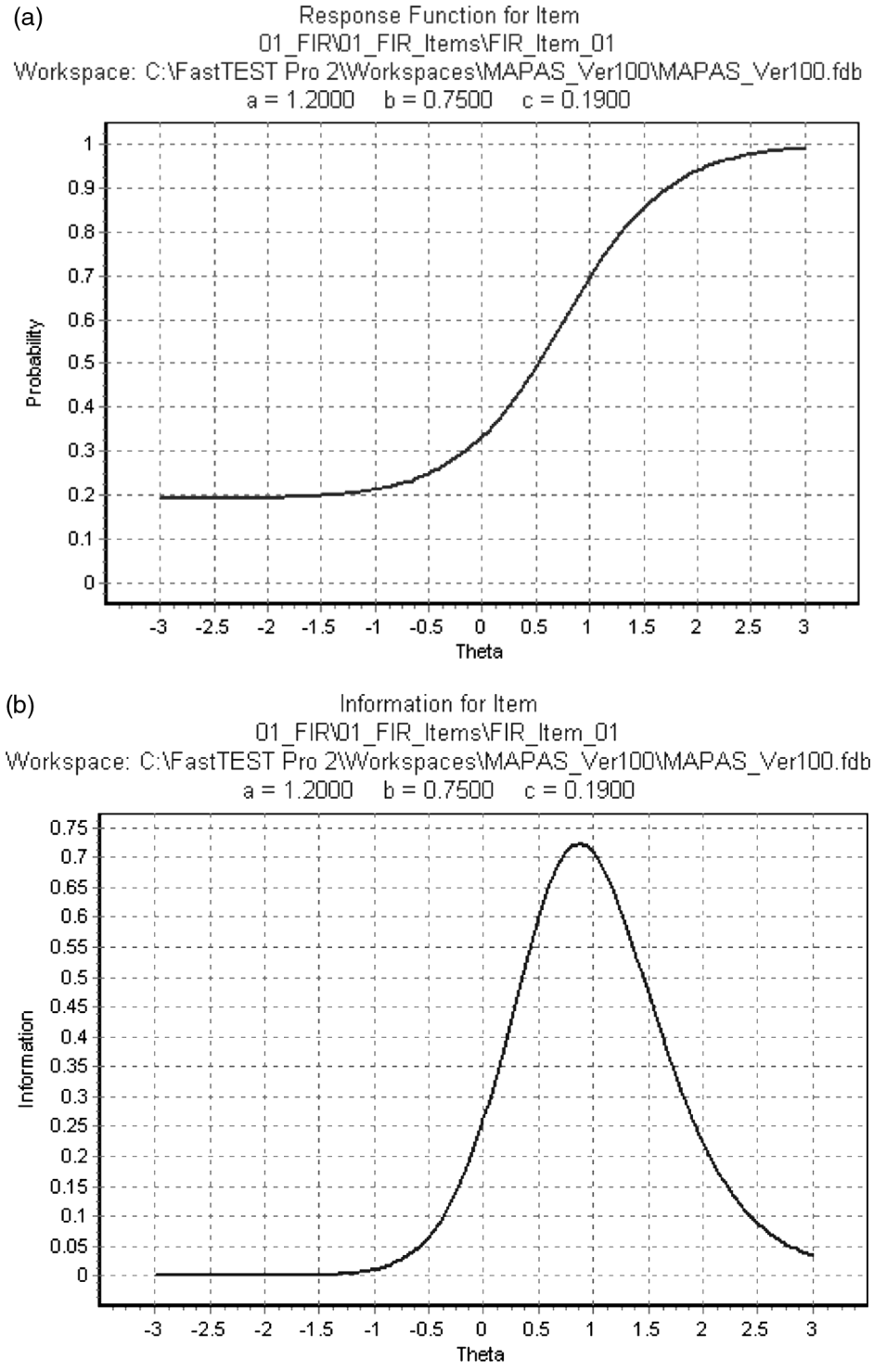

FIGURE 10.3. FastTEST item response theory (a) item response function and (b) item information function for an item. Copyright © 2006 Assessment Systems Corporation. Reproduced with permission.

type, keyed responses, author, source), Statistics (both IRT and classical), and Notes.

Thus, using multipart and multifield records in a database system, Windows item bankers replicated, automated, extended, and greatly improved the efficiency of the functions of the index cards originally used for test item storage and retrieval. In addition, however, computer-based database systems also permit highly efficient and virtually error-free search and retrieval. Windows item bankers capitalize on this capability to permit efficient and effective test assembly.

For simple test assembly, a test of a specific number of items can be randomly selected from an item bank or portions of an item bank. The latter approach would be used in successive cumulative searches to create a test that has a specific content structure with proportional representation of a larger content domain.

For constructing tests with deliberate nonrandom item selection, item bankers allow intelligent searching of information on items. Figure 10.4 shows the FastTEST item search window. As the figure shows, one can specify items in a bank, multiple banks, or portions of a bank to be searched. Searches can be implemented within most of the fields in the item record. Item identifiers, keywords, and item descriptions frequently include content or item-type information that is not included in the item bank structure, allowing item subsets to be identified that have specific content or structural characteristics (e.g., all free-response items, if that information is included in any of these fields). In addition, Figure 10.4 shows that separate or simultaneous searches can be made on all the psychometric data stored for each item. For classical test assembly, item–total correlations (discriminations) in a given range can be searched for while at the same time searching for items that have *p* values (difficulties) in a desired range. Searches of this type can be combined with content searches either by restricting the portion of the item bank searched to a particular content subsection of the item bank or by simultaneously limiting the statistics search within item subsets that match content search criteria.

When IRT item statistics or parameters are available in the item records, item banks or portions thereof can be searched for various combinations and ranges of the IRT discrimination, difficulty, and pseudo-guessing parameters. For more sophisticated IRT test assembly, FastTEST allows searches on item information, thereby helping the test developer create tests with a desired test information function.

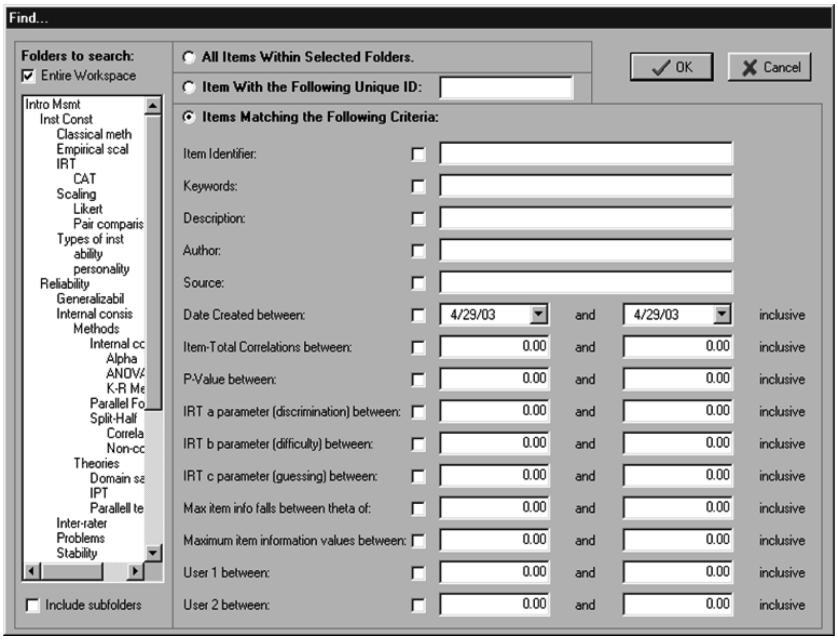

FIGURE 10.4. Item search options in FastTEST. Copyright © 2006 Assessment Systems Corporation. Reproduced with permission.

For example, a test developer might implement successive cumulative searches for items that have their maximum information values within specified ranges of the trait (theta) and for which the maximum information values are contained within a designated range. The result would be a set of items (if they existed in the bank) that had high maximum information throughout the theta range of the combined searches. In all cases, item bank searches frequently occur in a second or two, with slightly longer times for very large banks.

The result of an item search in an item banker is typically a list of items that meet the search criteria. Given that subset of items (akin to looking through the card file drawer and selecting a tentative set of items), the test developer will then usually select the items to include in the final test. This selection can be done in several ways. One approach is simply to randomly select a subset of items from among the items that meet the search criteria. A second is to browse through the item text and other information on the items and manually select items from the searched pool of item candidates. In either case, items are added to the test with a simple click of a mouse.

Once the items that will make up the test are selected, the next step is frequently to reorder the items in the test as desired. In a Windows item banker, this can be done by dragging and dropping within the item list that makes up the test, if a defined order is desired, or by randomly scrambling the items. If alternate forms of the test are needed, the test constructor can create any number of randomly scrambled alternate forms with a click of a mouse.

Before a test is finalized, a test developer might want to examine the statistical characteristics of the test on the basis of item statistics in the item bank. A few mouse clicks will make this information available in either graphical or tabular form. Figure 10.5 shows a frequency distribution of classical item difficulties in a test assembled with FastTEST; a similar graphic is available for item discriminations. If the test developer is not satisfied with the statistical characteristics of the test (before it is administered), she or he can drag and drop additional items and instantly reexamine the revised test's statistical characteristics.

If the test has been constructed with items for which IRT item parameters are available, the banker can display a test information function (e.g., Figure 10.6), a test response function, or a test standard error function. The test information in Figure 10.6 shows that the test being assembled provides a considerable amount of information around $\theta = 1.3$

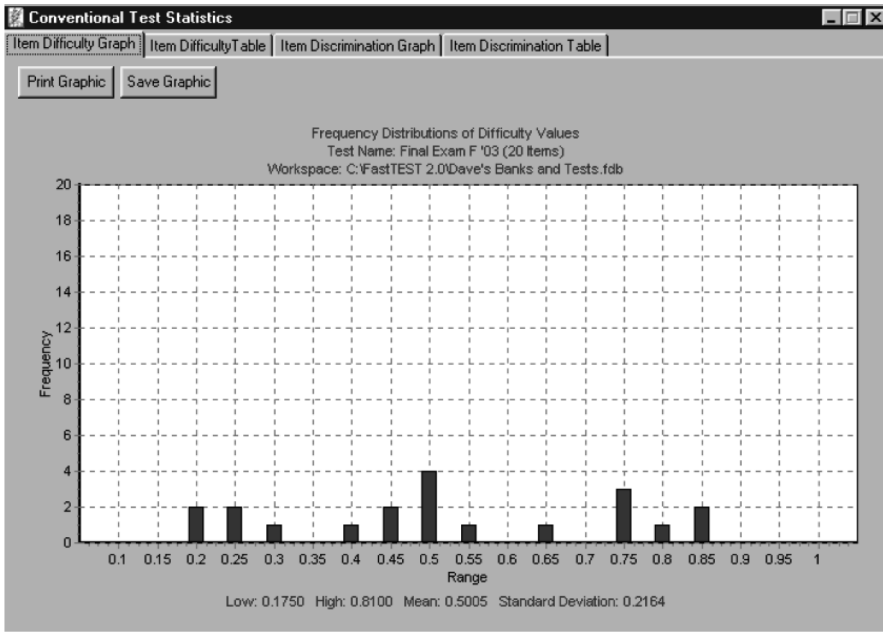

FIGURE 10.5. Frequency distribution of proportion correct for a 20-item test.
Copyright © 2006 Assessment Systems Corporation. Reproduced with permission.

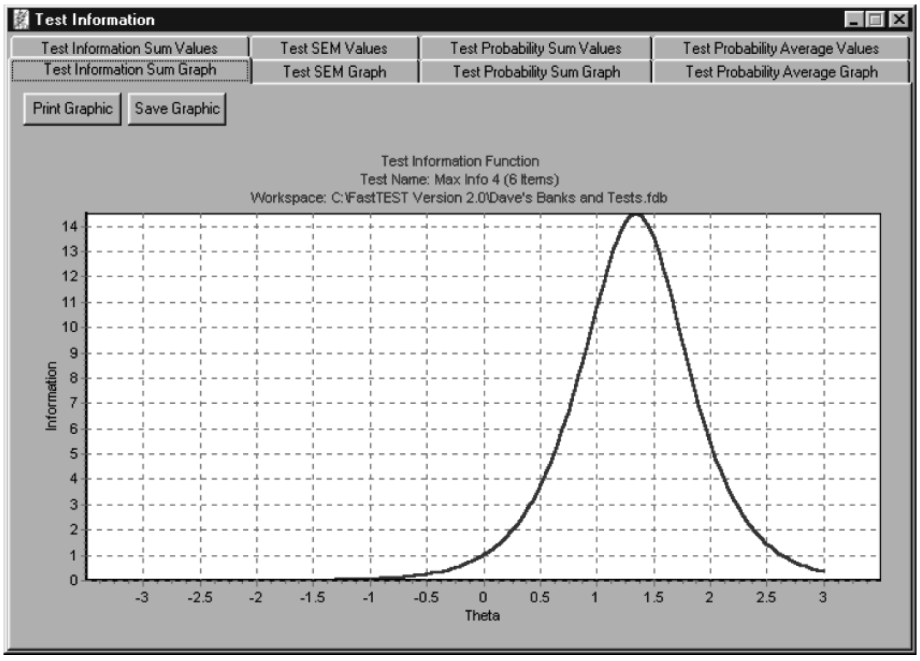

FIGURE 10.6. A test information function. Copyright © 2006 Assessment Systems Corporation. Reproduced with permission.

and very little information elsewhere along the theta scale. It is, as designed, a good test for differentiating individuals who are below or above $\theta = 1.3$ but has little measurement precision outside a range of about ± 1 standard deviation around that point. This test provides virtually no precision for theta values below average (0.0). Depending on the purpose for which the test was being built, revision of this test might be in order before it is used.

The final phase of item banking and test assembly is frequently one or more printed tests. Typically, item bankers will permit the insertion of instructions into the test document before it is printed. Most item bankers will also output a printed test with final or near-final formatting. Some will output the test as a rich text file that then can be further formatted in a word processing program before printing. They will also typically output a scoring key for each form of the test that they print. Of course, if no changes have been made in the items when they are formatted as the final test or when they are printed by the banker, no proofreading of item text is required.

The final component of the test development cycle is updating the item bank with item statistics from item analyses of the data from the P&P-administered test. This updating can be done manually, with appropriate item statistics typed into the item record for each item in the bank, or, more efficiently and accurately, item statistics output from item and test analysis software can automatically be imported into the item banker. For example, FastTEST includes a wizard that will import item statistics output from any item analysis software, thus completing the test development cycle.

Thus, the marriage of computer technology and database software designed specifically for testing has, in a short period in the history of testing, radically changed the way in which tests can be developed. The key element is banking software that allows the user to create structured banks, search the banks on a wide range of criteria, and assemble tests on the basis of both psychometric and content considerations. The process of creating a test has transitioned from a tedious and error-prone process that consumed many person hours to a simple process that can occur in a matter of minutes, once one or more properly constructed banks of test items have been entered into a well-developed item-banking system.

Electronic Test Delivery

The major change in how tests are delivered was also a result of the introduction of the PC and, for item

bankers, with additional impetus from the availability of Windows software. Electronic testing, or computer-based testing (CBT), began in the early 1970s. CBT eliminates both printed tests and answer sheets—test questions are stored in the computer and displayed on a monitor, and answers are generally entered by keyboard and more recently by mouse. Early CBTs were delivered on mainframe time-shared computers (De Witt & Weiss, 1974). These were typically connected by dial-up telephone modems operating at 10 or 30 characters per second connected to "dumb" character-based displays. It quickly became apparent, though, that this computer configuration was inadequate for test delivery. In addition to being limited to test items that were entirely character based, transmission and display time were far too slow and the system response time of these early systems was far too unpredictable; processing of a single item response and transmission of the next item sometimes took 30 seconds or more.

In the mid-1970s, minicomputers became available for testing research, and I used them for early delivery of adaptive tests (e.g., De Witt & Weiss, 1976). Because these computers were dedicated to the single task of testing and monitors were hardwired to the computer, system response time and display time were virtually instantaneous. They were, however, also limited to solely character-based test items. These systems, however, foreshadowed the primary improvements to be realized from CBT: (a) a fixed set of items could be administered in different orders to different examinees, (b) different subsets of items could be administered to different examinees to achieve certain measurement objectives, (c) item response data were instantly captured and cumulated across examinees and easily prepared for analysis, and (d) tests were immediately scored and individual reports could be prepared and available in seconds.

Randomized tests. Randomized P&P test forms have sometimes been used in large testing programs to minimize copying among examinees in adjacent seats. In this application, two or three versions of a test are created with the base form randomized once or twice to create alternate forms. In CBT, the process of whole-test randomization can be extended to separate randomizations of item order for each examinee. This randomization can be useful in CBT environments in which a number of examinees are taking the same test in the same computer lab, to minimize answer copying by students whose visual field might include another examinee's monitor.

A second form of CBT individualized randomization involves randomly selecting a subset of items from a larger domain of items. For example, an item bank might contain 200 items that define a specific content domain, and any given examinee might receive 50 items randomly selected from that domain. This process results in a relatively unique set of items administered to each examinee (there will, of course, be random item overlap among examinees) and a random sequence of items administered to each examinee. A variation of random item selection uses a stratified approach to randomly select items from a domain that has been subdivided into subdomains. For example, a mathematics domain might be stratified by type of operation—addition, subtraction, multiplication, and division. A randomized CBT might be designed to administer 10 items randomly selected from each subdomain to each examinee, for a test consisting of 40 items. Both whole-test randomization and subdomain randomization can be implemented with most PC-based testing systems (e.g., the FastTEST Professional Testing System; Assessment Systems Corporation, 2008) as well as Web-based testing systems (e.g., FastTEST Web; http://www.fasttestweb.com).

Random item selection thus explicitly implements the concept of domain sampling, commonly articulated as the basis for reliability theory using classical test development methods, and minimizes answer copying in a CBT environment. The process is, however, contradictory to the classical process of constructing some tests. In the first 60 or so years of P&P testing, some tests were (and still are) built with items in increasing order of difficulty, on the assumption that examinees perform better when they have a sufficient number of easy items at the start of the test to reduce test anxiety. Obviously, either strictly or stratified randomized CBTs cannot easily accommodate this rationale. Little to no research has addressed the effects of item randomization on examinees and their test scores as compared with tests built to accommodate warm-up effects.

Intelligent item selection. Contrasting with randomized item selection in CBT are tests that use intelligent item selection. These CBTs fall into three major types: linear-on-the-fly tests, sequential tests, and adaptive tests, each designed to implement different measurement objectives.

Linear-on-the-fly tests. Linear-on-the-fly tests are essentially fixed-length randomly selected tests with constraints (Thompson, 2008). They operate from a large item bank with IRT parameters available for each item. Items are pseudo-randomly selected, but the IRT parameters are used as the test is delivered to each examinee to monitor the psychometric characteristics of the test in real time, and the results are compared with psychometric targets defined in advance. As a result, tests for each examinee will have similar psychometric characteristics, but they will be achieved using different subsets of items for each examinee. A major advantage is that of equalizing item exposure to increase the security of an item bank across tests that are administered over time to a large group of examinees.

Sequential tests. Sequential tests are typically designed to make classifications. These tests might be used in a school to make pass–fail decisions, in an employment context to make a decision to hire or not hire, or in a professional certification program to determine whether an individual meets specified certification criteria. Although some sequential tests use random item selection, the more effective tests use intelligent item selection to the extent that psychometric information on test items is used to order items before item delivery. Then, given the fixed item order, items are administered and scored one at a time. After each item is administered a classification algorithm, such as the sequential probability ratio test (Eggen, 1999; Reckase, 1983), is used to attempt to make a classification of the examinee. If a classification can be made within prespecified error tolerances, test administration is terminated for that examinee. If a high-confidence classification cannot be made, the next item is administered and the decision criteria are again reevaluated. The result is tests that can make accurate classifications very efficiently, with a minimum number of items for each examinee.

Adaptive tests. Computerized adaptive tests (CATs) implement fully intelligent item selection (Wainer, 2000; Weiss, 1985, 2004). Unlike sequential tests that use a fixed order of items and allow only test length to vary, the more advanced versions of CATs also allow each examinee to start his or her test with different items and to receive quite different sets of test items.

Several varieties of CATs exist; to some degree, they all dynamically select items to be administered to each examinee on the basis of the examinee's answers to previous items in the test. Some use prestructured item banks in which an examinee's next item is determined by a branching tree structure in which a correct answer to a given item results in a particular next item and an incorrect answer leads to a different item. Others divide items into subsets or "testlets" (Mead, 2006; Wainer, Bradlow, & Du, 2000; Weiss, 1974). In this approach, each testlet, or minitest, is scored, and on the basis of that score a decision is made as to which testlet is to be administered next. In yet another approach, test items are stratified by item difficulty (Weiss, 1973) or discrimination (Chang & van der Linden, 2003) and administered sequentially within or between strata.

The most flexible and, therefore, efficient CATs are the fully adaptive CATs based on IRT. These CATs are based on IRT item information functions (e.g., Figure 10.3b), which are transformations of the IRT item parameters. The use of information functions allows each examinee to start the test with a different item if valid prior information is available. Then, on the basis of the answer to that item, a score is computed for that examinee, expressed on the IRT trait scale (theta) using estimation methods that take into account which answer the examinee gave to the item (correct or incorrect, keyed or not keyed, or which rating scale alternative was selected) and the item parameters for that item. The updated score is then used to select the one unadministered item out of an entire bank that provides the most information for that examinee, which is also the item that maximally reduces the uncertainly associated with the theta estimate (as expressed in the individualized standard error of measurement associated with the theta estimate). One or more termination criteria are then consulted—these criteria are typically a specified minimum value of the standard error or some maximum number of items. If

the examinee has not met one of the termination criteria, the current theta estimate is used to select the next best item, and the process continues. When a termination criterion is met, the test is ended and the final theta estimate and its standard error are recorded for that examinee.

CAT was first implemented primarily in the ability–achievement domain (Weiss & Betz, 1973). In recent years, it has begun to be used in personality measurement (Reise & Henson, 2000) and in medical research by measuring patient-reported outcomes of medical processes and procedures (Reeve et al., 2007). Early CAT research in the ability–achievement domain (e.g., Kingsbury & Weiss, 1983; McBride & Martin, 1983) indicated that CATs could measure with precision equal to that of conventional tests using at least 50% fewer items; these findings have been supported and extended in numerous applications (e.g., Mårdberg & Carlstedt, 1998; Moreno & Segall, 1997). More recent research in the personality and mental health domains has indicated that reductions in test length as high as 95% can be obtained on a general impairment scale and as high as 85% for measuring four subscales, with little or no reduction in measurement accuracy from full-length tests that use an entire large item bank (Gibbons et al., 2008).

The major advantage of CAT is the ability to design and deliver efficient tests that measure all examinees with an equal level of precision. This means that in a CAT properly designed for this measurement objective, all examinees will be measured with the same standard error and minimum test length, an objective not easily achieved with any other kind of test. Obviously, because of the extensive real-time calculations necessary to implement CATs, they cannot be delivered by any other means than computers. The FastTEST Professional Testing System (Assessment Systems Corporation, 2008), in conjunction with CATSim (Assessment Systems Corporation, 2010a) permit the design and delivery of fully adaptive tests using IRT given an item bank of items with estimated IRT parameters.

Other advantages of CBT. Clearly CBT has changed the way tests are delivered. Only rarely now do different individuals receive the same fixed set of items in the same order. CBT also has changed the way test data are captured, stored, and used as well as allowing completely new kinds of tests.

Data capture. Because all forms of CBT involve electronic item delivery and the immediate electronic capture of item responses, all of the problems associated with printing and distributing test booklets and answer sheets, as well as the unreliability of the scanning process, have disappeared. Item response data in CBT are stored as each examinee answers each item and can be accumulated across examinees with ease. If the test is randomized, sequential, or adaptive, responses are automatically reordered into a common order to allow analysis. Depending on the software system, cumulated item responses can be immediately analyzed, and the results are available at any time, even on a real-time basis if desired.

An additional potential advantage of CBT is the availability of item response times. The PC can record the time from when the item is presented to the examinee to when the examinee clicks the mouse to select an answer, clicks the "next" button to move to the next question, or both. Although such item response times can be somewhat unreliable, careful analysis of them might result in additional information, beyond the correctness or incorrectness of an examinee's answers, to assist in obtaining better measurements of the examinee's ability, attitude, or personality variables (Ferrando & Lorenzo-Seva, 2007).

Instant reporting. In addition to instant capture of item responses and instant scoring, CBT provides the capability to generate a wide variety of reports that can be displayed immediately to the examinee on completion of the test session (which can include multiple tests) or to a testing room supervisor, proctor, or teacher or that can be printed or saved in electronic files for later use. These reports can be as simple as a certificate of completion of the test with a passing score or as complex as a graphic plot of multiple test scores followed by a multipage interpretation of test results. Obviously, the combination of instant data capture and instant reporting permit test data and test results to be used for applied purposes far more quickly than was possible with P&P tests.

New types of measurements. A final major advantage of CBT is the capability to measure variables that cannot be easily measured with P&P. This capability includes the use of detailed color graphics in test items, audio and video, and animation. Audio and video are especially useful in certain types of language testing in which language segments are spoken through headsets and presented to examinees for translation and other processing. Other language-related applications include presentation of test items with an audio or video option for examinees who have reading limitations. More recently, innovative item types have expanded on simple multimedia to include interactive simulations in an attempt to provide more fidelity to the measurement of complex processes.

CBT also allows the measurement of some abilities and other variables for which P&P tests are not optimal. For example, although memory is an important ability for success in academic environments and many jobs, there have been no major P&P tests of memory ability because measuring memory requires an interactive, individualized, and controlled process that would be very labor intensive. Such tests, however, could easily be developed in a CBT environment (e.g., Letz, 2003) in which it is possible to control the period of time that material is displayed, to use a wide variety of material—words, phrases, audio clips, and video clips—and to test for recall after specified time intervals. The process can also be made adaptive in that display and recall times could be individualized for an examinee on the basis of his or her performance on earlier tasks.

Another type of new item that is uniquely computer administered is the so-called scenario or problem-solving item. In this kind of test, a situation is described and the examinee is given a choice of various elements of information that pertain to the situation. After the examinee consults the selected information, questions are posed to him or her that then lead to other information sources. The process continues until some resolution is reached, which, depending on the sequence of choices made by the examinee, could result in an adequate solution to the original problem or to solutions that are inadequate to various degrees (and, therefore, result in lower scores). This kind of interactive problem-solving test is most notable in the medical training (e.g., Dieckmann, Lippert, Glavin, & Rall, 2010) and licensing (e.g., http://www.nbme.org) environment in which the "patient" presented in the original scenario either is cured or dies or the sequence of choices made by an examinee results in some intermediate suboptimal state.

THE INTERNET'S IMPACT ON TESTING

As with many functions performed with PCs, the rise of the Internet and the World Wide Web began to affect the test development and delivery process beginning in the late 1990s, as it has affected many other areas of psychological research (e.g., Gosling & Johnson, 2010). Test development is frequently a process that draws on the expertise of a variety of personnel. In a large testing program, such as that of a school system or a licensing or certification organization, test items are written by a number of people with specific expertise, some of whom are geographically dispersed across a country or even different countries. Although it is possible to collect test items from remote experts by sending e-mail files to a central location for item bank development, the Internet presented an opportunity to allow test items to be entered into item banks from any computer with access to it. Thus, once an item bank is developed, software systems such as FastTEST Web (http://www.fasttestweb.com) allow item writers (with appropriate security safeguards) to access designated portions of item banks and to directly add new items to the banks.

A second stage of item bank development, also available for remote access in systems such as FastTEST Web, is item review and editing. Item reviewers and editors have different skill sets than item writers and are, therefore, likely to be different personnel and located in different places. FastTEST Web defines a different role for item editors and allows the test development supervisor to limit their activities to only items that are appropriate for review. Other roles include test assembly and test (vs. item) review; each activity can be done remotely by any number of appropriate personnel at any location without the need to send any material to a central location for processing. The result is an item and test development process that is even more

efficient than that possible using PC item bankers. Of course, Internet item bankers such as FastTEST Web include all the functionality in item banking and test assembly as the PC-based item bankers, plus the capability of running a wide range of reports on banks and tests from any location. In addition, the tests developed through Web-based bankers can be printed or delivered directly through the Web to examinees at any location.

A major characteristic of PC-based CBT is that the test and test data are stored on individual computers on which tests are administered or on a network server that is hardwired to the testing computers. When tests are delivered on stand-alone computers, test data must be collected from each computer and aggregated for further storage and analysis. Although this process is easily automated to a degree, it still requires physical transmission of test data by some means. In addition, when the tests themselves are stored on independent testing stations, they must be individually installed and their existence on testing station hard drives can create potential item security problems unless the tests are well encrypted.

Internet test delivery solves these problems, although not without creating some others. In Internet testing, which has become very popular in recent years, tests are stored centrally, along with all the information necessary to score them (e.g., IRT item parameters or classical item option score weights). Items are sent through the Web, one or more at a time, presented to the examinee, and the response is accepted and transmitted back to the server. The next item, or set of items, is selected and presented, and the process is continued until the test is completed. At the end of the test, as in PC-based testing, an assortment of reports is available for presentation to the examinee and other appropriate personnel. The advantages are, of course, that tests can be delivered to any computer that has Internet access, test items are not stored on the testing computer but rather appear only on the monitor screen, and all test data are instantly stored in a central database and are available for analysis at any time.

A number of new problems are raised by Internet testing, however (Naglieri et al., 2004). In PC-based testing, which has typically been implemented in testing centers or testing labs, monitors and other associated equipment can easily be standardized. Standardization involves specifying a defined set of conditions under which the measurements are obtained that is designed to control extraneous influences that might affect the measurements (and add error to them). Internet-delivered tests, however, are frequently administered to individuals using their own computers, which can be desktops, laptops, or notepads. These computers might have different display resolutions and different display sizes. As a consequence, the same test item might be rendered differently on different computers. In the P&P testing era, standardization of the test material was heavily emphasized—a given test was always formatted and printed in exactly the same way. With Internet testing, because of the various displays possible under certain remote test administration conditions, there is a danger that the characteristics of the display will degrade the standardization required for adequate measurement of some variables.

A second threat to standardization in Internet test delivery is that of the testing environment. P&P testing and most applications of PC-based testing emphasized that test delivery should occur in a quiet, well-controlled environment in which examinees could concentrate on the tasks and questions posed in the test with minimal distraction. Test instructions were standardized, lighting and room temperature were controlled, and other outside influences were eliminated or minimized. Unless an Internet-delivered test is administered in a space devoted to testing or a location that is under the supervision of a test administrator, there is no control over the testing environment. To the extent that test scores can be influenced by nonstandardized testing conditions—noise, other people present, variations in temperature and lighting, and a host of other factors that might exist in nonstandardized testing environments—scores from such tests cannot be relied on to be as precise and valid as those from tests taken under controlled conditions.

When Internet-based tests are delivered in unsupervised environments, it is also frequently not possible to know exactly who is taking the test or what they are doing during test delivery. There might be other people available to the examinee who are being consulted during test administration, or in an

extreme case someone other than the presumed examinee might complete the test or a portion of it. In addition, unless the test delivery software explicitly locks the examinee's computer from accessing its hard drive and simultaneously locks the Internet browser from accessing other Web sites, an examinee might access other electronic sources during the test to answer the test questions. Even under complete electronic lockout, an unsupervised examinee can access printed sources without the knowledge of the organization providing the test. It should thus be clear that unsupervised Internet test delivery is not appropriate for high-stakes tests that are being used to make important decision concerning an examinee.

A final source of lack of standardization of Internet-delivered tests lies in the nature of the Internet itself. The Internet is basically an extremely sophisticated time-sharing system, but one that involves a great number of loosely networked computers. As such, delays always exist between when information is sent to when the sending computer receives the information that it requests. These delays can be minimal—a second or two—or quite a bit longer, but they are always unpredictable. They result from a combination of many factors, including the amount of traffic on the Internet, the speed of transmission over the various components of the system used for a message to reach its destination server and return, the speed of the server and the load on it when the message is received, and server processing time.

For many testing purposes, these delays might be relatively inconsequential, especially because many people have become accustomed to them. However, the delays can accumulate in testing applications in which items are delivered one item at a time, such as sequential testing and adaptive testing. In these applications, in addition to the system delays, computations must be done between each item delivery, thus potentially exacerbating the delays. Delays of several seconds between items can result in a testing experience that is less than optimal for many examinees, and their unpredictability might be a source of test anxiety for some examinees.

CONCLUSIONS

The first 60 or so years of psychological, educational, and personnel testing were dominated by the P&P test. Item banking, test assembly, and test scoring were entirely manual procedures that were labor intensive, tedious, and prone to errors. Tests were highly standardized, as were conditions of administration. Changes began to occur with the introduction of electronic optical mark readers that reduced test scoring to a relatively accurate partially automated procedure that dominated standardized testing for many decades. The introduction of the PC in the mid-1980s, however, began a major evolution of testing away from the traditional way of building tests and delivering them.

The years since 1985 have seen computers automate the processes of item banking, test assembly, test analysis, and test delivery. The PC allowed the development of new modes of test delivery—random, sequential, and adaptive—and new kinds of test items. The advent of the Internet extended test delivery to any computer that could connect to it, albeit not without some problems.

The result of this evolution of testing is a set of processes that are considerably less labor intensive, more accurate, and more efficient. In the process of this ongoing conversion, numerous questions have arisen, some of which have not yet been satisfactorily studied or even addressed. Few research questions surround computerized item banking and test assembly. The major questions have risen in the context of CBT. In the early days of CBT, it was natural to address the question of whether CBTs functioned the same as P&P tests. Generally, it was found that they did (Mead & Drasgow, 1993), although early research indicated some differences on reading comprehension tests (e.g., Kiely, Zara, & Weiss, 1986; Mazzeo & Harvey, 1988). As CBTs begin to be used to measure constructs that cannot be measured by P&P, however, comparability is no longer an issue, and CBTs will have to be validated on their own merits.

Obviously, a host of questions exist about how to best implement CATs and sequential tests that have resulted in substantial research over the past 20 or 30 years and will continue to do so (for an extensive bibliography of CAT research, see http://iacat.org/biblio). CBTs also raise a number of questions about the psychological environment of testing that have generally not been addressed. In the process of

creating a test of appropriate difficulty for each examinee, CATs create a different psychological environment than do P&P tests. Does that difference affect examinee performance? PC-delivered tests have virtually no delays between items in comparison to Internet-delivered tests. Do the unpredictable delays in Internet-based testing affect examinees' test anxiety and thereby influence test performance? Do the random variations in the testing environment that occur for unsupervised Internet-based testing affect test scores?

Finally, as the Internet continues to pervade people's activities through various electronic devices, some have suggested that certain kinds of psychological measurements (e.g., attitudes, personality variables) can be delivered by portable electronic devices such as PDAs and cellular phones. If these modes of test delivery are implemented, the usefulness of the resulting measurements will have to be carefully scrutinized because of the extremely variable testing conditions under which such measurements will be obtained. As the American Psychological Association Task Force on Psychological Testing on the Internet (Naglieri et al., 2004) concluded,

> Despite the flash and sparkle of Internet testing, critical questions of the validity of the inferences made from test scores must be demonstrated. This is a fundamental issue of test validity that must be weighed in relation to the ease of availability, cost, and convenience of Internet testing. (p. 161)

References

Assessment Systems Corporation. (1987). *User's manual for the MicroCAT Testing System* (2nd ed.). St. Paul, MN: Author.

Assessment Systems Corporation. (2008). *Manual for the FastTEST Professional Testing System, Version 2.* St. Paul MN: Author.

Assessment Systems Corporation. (2010a). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Author.

Assessment Systems Corporation. (2010b). *User's manual for the FastTEST 2.0 Item Banking and Test Development System.* St. Paul MN: Author.

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement, 27*, 262–274. doi:10.1177/0146621603027004002

De Witt, J. J., & Weiss, D. J. (1974). *A computer software system for adaptive ability measurement* (Research Report 74-1). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Dewitt, L. J., & Weiss, D. J. (1976). Hardware and software evolution of an adaptive ability measurement system. *Behavior Research Methods and Instrumentation, 8*, 104–107. doi:10.3758/BF03201754

Dieckmann, P., Lippert, A., Glavin, R., & Rall, M. (2010). When things do not go as expected: Scenario life savers. *Simulation in Healthcare, 5*, 219–225. doi:10.1097/SIH.0b013e3181e77f74

DuBois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn & Bacon.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249–261. doi:10.1177/01466219922031365

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525–543. doi:10.1177/0146621606295197

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361–368. doi:10.1176/appi.ps.59.4.361

Gosling, S. D., & Johnson, J. A. (Eds.). (2010). *Advanced methods for conducting online research.* Washington, DC: American Psychological Association. doi:10.1037/12076-000

Kiely, G. L., Zara, A. R., & Weiss, D. J. (1986). *Equivalence of computer and paper-and-pencil Armed Services Vocational Aptitude Battery tests* (AFHRL-TP-86–13). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York, NY: Academic Press.

Letz, R. (2003). Continuing challenges for computer-based neuropsychological tests. *Neurotoxicology, 24*, 479–489. doi:10.1016/S0161-813X(03)00047-0

Mårdberg, B., & Carlstedt, B. (1998). Swedish Enlistment Battery: Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB.

International Journal of Selection and Assessment, 6, 107–114. doi:10.1111/1468-2389.00079

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Report No. 88-8; ETS RR No. 88-21). New York, NY: College Entrance Examination Board.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York, NY: Academic Press.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*, 185–187. doi:10.1207/s15324818ame1903_1

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458. doi:10.1037/0033-2909.114.3.449

Moreno, K. E., & Segall, O. D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169–180). Washington, DC: American Psychological Association. doi:10.1037/10244-018

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59*, 150–162. doi:10.1037/0003-066X.59.3.150

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). New York, NY: Academic Press.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45* (5, Suppl. 1), S22–S31.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347–364. doi:10.1177/107319110000700404

Thompson, N. A. (2008). A proposed framework of test administration methods. *Journal of Applied Testing Technology, 9*(5). Retrieved from http://www.testpublishers.org/mc/page.do?sitePageId=112031&orgId=atpu

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1996). *ConTEST 2.0: A decision support system for item banking and optimal test assembly* [Computer program and manual]. Groningen, the Netherlands: Iec ProGAMMA.

Vale, C. D. (2006). Computerized item banking. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 261–285). Mahwah, NJ: Erlbaum.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Norwell, MA: Kluwer.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774–789. doi:10.1037/0022-006X.53.6.774

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*, 70–84.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive?* (Research Report 73-1). University of Minnesota, Department of Psychology, Psychometric Methods Program.

CHAPTER 11

SCALING, NORMING, AND EQUATING

Michael J. Kolen and Amy B. Hendrickson

Test scoring typically begins with scores on individual test items, which are referred to as *item scores*. Item scores can be incorrect or correct, can involve multiple score points such as when human scorers score an essay on a 5-point rubric, or can indicate an examinee's level of agreement with an idea. The *raw score* for an examinee on a test is a function of the item scores for that examinee. Raw scores can be as simple as a sum of the item scores or be so complicated that they depend on the entire pattern of item responses. Raw scores are typically transformed into scale scores using a process referred to as *scaling* so as to facilitate score interpretation. Such scores are reported on a score scale.

Norming involves collecting data from a norm group of individuals to produce norms. To facilitate score interpretation, an individual's scores can be compared with scores for the norm group to assess the individual's relative standing in the norm group. The usefulness of the norms depends on how significant the norm group is for the score interpretation to be made. Incorporating information from a norming study into the score scale is one way to improve score interpretability.

Alternate test forms are often used with educational and some psychological tests for reasons of test security and so that examinees can be tested more than once. Because it is impossible for test developers to build alternate test forms that are of equal difficulty, test form equating methods are used to provide statistical adjustments to scores so that scores from the alternate forms can be used interchangeably. After the equating process, scores from the alternate forms are reported as scale scores. Reporting scale scores on the alternate test forms makes it more likely that the same reported score on two alternate forms is indicative of the same level of the construct being tested.

Procedures for scoring, scaling, norming, and equating can work together to facilitate the usefulness of reported scores. Such procedures can facilitate the proper use of test scores in making important psychological and educational decisions. Scoring, scaling, norming, and equating procedures used in concert allow test users to identify trends in test results over time and to ensure that educational and psychological criteria have the same meaning over time. Note that the terms *psychological scale* and *clinical scale* are often used in the literature to refer to groupings of items that measure a similar construct, such as the Neuroticism Scale on the revised NEO Personality Inventory (McCrae & Costa, 2010). However, in this chapter the term *scale* is used, as defined earlier, to refer to the scale used to report scores.

Theoretical and practical considerations for scoring, scaling, norming, and equating are described in this chapter. For more detailed treatments of certain aspects of these areas and numerical examples, refer to Angoff (1971); Embretson and Reise (2000); Dorans, Pommerich, and Holland (2007); Flanagan (1951); Holland and Dorans (2006); Kolen (2006); Kolen and Brennan (2004); Petersen, Kolen, and Hoover (1989); and von Davier (2011). The American Educational Research Association, the American Psychological Association, and the National

DOI: 10.1037/14047-011
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

Council on Measurement in Education (1999) jointly provided standards that should be used for scaling, norming, and equating in practice (see Chapter 13, this volume).

This chapter begins with a discussion of different perspectives on scales followed by a description of score scales for individual tests, scales for batteries and composites, and vertical scales. The discussion of scales is followed by discussions of norms and equating.

SCALING PERSPECTIVES

A variety of score scales and approaches to constructing them have been used in the development of scales for psychological and educational tests. These approaches depend on the nature of the construct being assessed and on the perspective taken by the scale's developer.

In one approach, a psychometric model is used to drive the development and scaling of tests. Thurstone (1925) developed one of the first such psychometric models, which led to a process of choosing items and assigning scale scores to individuals. Later, Thurstone (1928) made claims about the equality of units of measurement that arose from the application of his approach. Guttman (1944) developed a model for scaling attitude items and individuals on the same scale. His model included criteria to assess whether a scale could be constructed, and it focused on appropriately rank ordering examinees and placing individuals and items on the same scale. Rasch (1960) models also place individuals and items on the same scale and have been used to develop scales for a variety of different types of psychological constructs (Wright & Stone, 1979). Wright (1977) summarized the psychometric model–based approach to developing scales for tests as follows:

> When a person tries to answer a test item the situation is potentially complicated. Many forces might influence the outcome—too many to be named in a workable theory of the person's response. To arrive at a workable position, one must invent a simple conception of what we are willing to suppose happens, do our best to write items and test persons so that their interaction is governed by this conception, and then impose its statistical consequences upon the data to see if the invention can be made useful. (p. 97)

Thus, with the psychometric model–based approach, the focus of test development and scaling is on fitting the psychometric model.

Stevens (1951) classified scales as being nominal, ordinal, interval, or ratio. Suppes and Zinnes (1963) further developed this classification scheme, and Coombs, Dawes, and Tversky (1970, pp. 7–19) provided a summary. Coombs et al. pointed out that this theory requires that the relationship among individuals and the attribute be clearly defined and that there are many more scale types than the four proposed by Stevens (1951). In a discussion of the scaling of intelligence tests, which also applies to tests of other psychological constructs, Coombs et al. stated that because "no measurement theory [of this type] for intelligence is available . . . no meaning [from the perspective of this measurement theory] can be given" (p. 17) to the scores from intelligence tests. On the basis of this line of reasoning, applications of psychometric models such as those of Thurstone (1925, 1928) or Rasch (1960) are insufficient to support claims about properties of the scales that are developed. Similar points were made by Angoff (1971, pp. 510–511) and Yen (1986, p. 314). Recently, Michell (2008) called for the properties of scales to be taken much more seriously, although Kane (2008) expressed concerns about such an emphasis.

The development of scales for psychological and educational tests often focus on practical considerations that are consistent with the perspective of Petersen et al. (1989), who stated that "the main purpose of scaling is to aid users in interpreting test results. In this vein, we stress the importance of incorporating useful meaning into score scales as a primary means of enhancing score interpretation" (p. 222). This practical perspective, in which the primary purpose of score scale development is viewed as facilitating the interpretation of test scores by test users, is adopted in this chapter.

SCALES FOR A SINGLE TEST

A variety of scores and scales are used with educational and psychological tests. In this section, scores are discussed, followed by different types of scale-score transformations. Then different methods for incorporating meaning into score scales are considered, including normative and content meaning.

Scores

Kolen (2006) distinguished unit scores from item scores. A *unit score* is the score on the smallest unit for which a score is found, which is referred to as a *scoreable unit*. An *item score* is a score over all scoreable units for an item.

For multiple-choice or true–false test questions that are scored as either incorrect (0) or correct (1), such as on cognitive assessments of the National Assessment of Educational Progress (NAEP; Institute of Education Sciences, 2011) and the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2008), unit scores and item scores are often the same. For multiple-choice or true–false questions on noncognitive assessments such as the revised NEO Personality Inventory (McCrae & Costa, 2010) and the Learning and Study Strategies Inventory (Weinstein & Palmer, 2002), the particular item options endorsed form a test-taker profile and contribute to the test taker's score on various psychological scales. In these situations, the item scores vary depending on the psychological scale and are thus not meaningful on their own. Scale scores are the most basic score of interest for these assessments.

Raw scores are functions of item scores. The summed score is an often-used raw score, and it is calculated by summing the item scores. For cognitive tests that are scored as correct or incorrect, the summed score is the number of items the test taker answers correctly. For noncognitive tests, the summed (or scale) score is often the number of options or items endorsed that contribute to that particular psychological scale (e.g., the Extraversion scale on the revised NEO Personality Inventory; McCrae & Costa, 2010). Sometimes test developers decide to differentially weight the item scores using positive weights that produce weighted summed scores. With item response theory (IRT), proficiency estimates are often complex functions of the item scores (Embretson & Reise, 2000). These proficiency estimates can be viewed as raw scores.

Transformation of Raw Scores to Scale Scores

As Kolen (2006) indicated, raw scores have limitations as primary score scales for tests. Raw scores are dependent on the particular items on a test, and so they cannot be meaningfully compared when test takers take different test forms. In addition, raw scores do not carry normative meaning and are difficult to relate to meaningful generalizations to a content or psychological domain. For these reasons, raw scores are transformed to scale scores. Linear or nonlinear transformations of raw scores are used to produce scale scores that can be meaningfully interpreted. Normative and content information can be incorporated. Procedures for incorporating these types of meaning are considered next.

Incorporating Normative Information

Incorporating normative information begins with the administration of the test to a norm group. Statistical characteristics of the scale-score distribution are set relative to this norm group. The scale scores are meaningful to the extent that the norm group is central to score interpretation (Kolen, 2006).

For example, the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1989) was administered to a national norm group of nonpatient subjects intended to be representative of adults in the United States. These data were used to establish linear *T* scores with a mean of 50 and standard deviation of 10. By knowing the mean and standard deviation of the scale scores, test users are able to quickly ascertain, for example, that a test taker with a *T* score of 60 on the Depression scale is 1 standard deviation above the mean. This information is relevant on the basis of the representative sampling of the norm group. Kolen (2006, pp. 163–164) provided equations for linearly transforming raw scores to scale scores with a particular mean and standard deviation.

Nonlinear transformations are also used to develop score scales. Normalized scores involve one such transformation. To normalize scores, percentile

ranks of raw scores are found and then transformed using an inverse normal transformation. These normalized scores are then linearly transformed to have a desired mean and standard deviation. Normalized scale scores can be used by test users to quickly ascertain the percentile rank of a test taker's score, using facts about the normal distribution. For example, the scale scores for all six psychological scales of the WAIS (Wechsler, 2008) are smoothed normalized scores set to have a mean of 100 and a standard deviation of 15. Thus, a score of 115 on the Perceptual Reasoning scale, for example, is 1 standard deviation above the mean and represents a percentile rank of approximately 84. Kolen (2006, pp. 164–165) provided a detailed description of the process of score normalization. Note that the authors of the Minnesota Multiphasic Personality Inventory have warned against using normalized *T* scores because they result in psychological profiles that are quite different from those based on linear or uniform *T* scores, because of the non-normal distribution of scores in the norm group (Hathaway & McKinley, 1989; Hsu, 1984).

Incorporating Content Information

Ebel (1962) stated, "To be meaningful any test scores must be related to test content as well as to the scores of other examinees" (p. 18). Recently, focus has been on providing content-meaningful scale scores.

For noncognitive assessments, such as clinical and personality inventories, meaning is often attached to the raw scores by assigning names to these psychological scales, such as *Anxiety* or *Attitude,* and a description of the behaviors and characteristics associated with these scales is provided. High scores on these psychological scales, then, easily give an indication of the test taker's personality or clinical state. Moreover, score ranges on the scales may be identified and given interpretive information. For example, interpretations of three *T*-score categories (≤ 44 = low, 45–55 = moderate, ≥ 56 = high) are provided for each of the scales of the revised NEO Personality Inventory (McCrae & Costa, 2010) that help to identify more specific behaviors and characteristics of test takers falling in these categories.

Another way to incorporate content information is to use scale anchoring, often used with cognitive assessments (Kolen, 2006, pp. 168). The first step in scale anchoring is to develop an item map that places each item on the score scale on the basis of where test takers have a particular probability of earning a particular score or higher on the item. Then a set of scale-score points is chosen, such as a selected set of percentiles. Subject matter experts review the items that map near each of the selected points and develop general statements that represent the skills of the test takers scoring at each point. See Allen, Carlson, and Zelenak (1999) for an example of scale anchoring with the NAEP; Zwick, Senturk, Wang, and Loomis (2001) for a study of alternative methods for scale anchoring; and American College Testing (2007) for an example of scale anchoring as used with its College Readiness Standards.

SCALES FOR BATTERIES AND COMPOSITES

Test batteries consist of tests in various content areas or items contributing to various psychological scales, with separate scores provided for each content area or scale. With test batteries, the processes of test construction and scaling are handled similarly for each test in the battery, making possible the assessment of test-taker strengths and weaknesses or profiles across test areas or scales. Sometimes composite scores are calculated, which are combinations of scores from some or all of the tests or scales in the battery. Using the same scaling procedures for each of the tests or scales in a battery facilitates the formation of such composites (Kolen, 2006).

Scale Comparability Across Tests in a Battery

When normative information is incorporated into the scale for a test, the same norm group is often used for all of the tests in the battery (Kolen, 2006). Using this normative information, the scale can be constructed so that the scale-score distributions for the tests in the battery are approximately the same for the norm group. For example, the same norm group and same linear transformation to a *T* score was used across all five of the revised NEO Personality

Inventory domains (McCrae & Costa, 2010). The *T* scores produce distributions that are very similar from domain to domain, with approximately equal percentages of scores at each *T*-score level. The *T* scores allow for percentile comparisons across the domains. Consider a test taker scoring 50 on the Neuroticism domain and 60 on the Agreeableness domain. Because of the *T*-score scale-score property built into the domains, this test taker's score is near the 50th percentile on the Neuroticism domain and near the 84th percentile on the Agreeableness domain. Relative to the norm group, the test taker exhibits more agreeable behaviors than neurotic behaviors.

Composites

Composite scores that reflect performance on two or more tests are often used (Kolen, 2006). Composite scores are typically a linear combination of scale scores on different tests. For example, six composite scores are reported from the WAIS (Wechsler, 2008), including the Full Scale IQ scale that is a measure of overall cognitive ability and is derived from the other five scales. Each is based on a sum of scale scores. Scores on the tests that are used to form composite scores are typically correlated. Effective weights (see Kolen, 2006), which are proportional to the correlation between the score on one of the tests making up the composite and the composite score, are sometimes used to index the contribution of each test to the composite.

VERTICAL SCALING AND DEVELOPMENTAL SCORE SCALES

Assessing the extent to which the aptitude or achievement of test takers grows from one year to the next is important for many cognitive applications. Growth might be assessed by administering alternate forms of the same test each year and charting growth as measured by changes in test scores from year to year and over multiyear periods. Apart from showing growth, when an assessment is to be administered across a wide range of age groups, using a single set of questions for all ages may be problematic. Most test takers would be measured imprecisely because the test would not be targeted at their current age or level. Younger test takers would be overwhelmed when presented with tasks that are much too difficult. Older students might be careless or inattentive when presented with many test questions that are too easy.

To address these issues, educational and psychological batteries are typically constructed using multiple test levels, in which each level is constructed to be appropriate for test takers at a particular grade or age. Vertical scaling procedures are used to relate scores on these multiple test levels to a developmental score scale that can be used to assess test-taker growth over a range of levels.

In this section are discussed the types of domains that are measured with vertical scales, different definitions of growth, designs for collecting data for vertical scaling, and statistical procedures used to conduct vertical scaling.

Structure of Batteries

Vertical scaling procedures are used with cognitive assessments such as aptitude test batteries (e.g., Cognitive Abilities Test; Lohman & Hagen, 2002), intelligence test batteries (e.g., WAIS; Wechsler, 2008), and achievement test batteries (e.g., Iowa Tests of Basic Skills; Hoover, Dunbar, & Frisbie, 2003). These types of batteries typically contain tests in a number of areas and are used with test takers in a range of ages. Students are administered test questions that assess content, skills, abilities, or aptitude relevant to that level. Going from early to later levels, the test questions become more difficult and the content becomes more advanced.

For many such batteries, test questions overlap from one test level to the next. Overlap reduces the development burden because the same items are used in adjacent test levels.

Designs for Data Collection and Statistical Methods for Vertical Scaling

Data are collected to conduct vertical scaling. Three of the most common designs used for data collection are the common-item design, the equivalent-groups design, and the scaling test design that were described by Kolen and Brennan (2004, pp. 377–380). The common-item design takes advantage of an overlap of test questions over levels. Each test level

is administered to students at the appropriate grade or age, and performance on the items that are common to adjacent test levels is used to indicate the average amount of growth that occurs from one grade or age to the next. The data from this design are used to place scores from all test levels on a common scale.

In the equivalent-groups design, randomly equivalent groups of test takers are administered the level appropriate for their grade or age and the level below their grade or age. By chaining across grades or ages, the data from this administration are used to place scores from all test levels onto a base level. The common-item and equivalent-groups designs are similar to equating designs, discussed later in this chapter.

In the scaling test design, a special test is constructed that spans the content across all of the grade or age levels of interest. Students in all grades or ages are administered the same scaling test. All students take the scaling test and the items appropriate for their level. The score scale is defined using scores on the scaling test. Scores on each test level are linked to the scaling test.

After the test is constructed and data are collected, psychometric methods, such as IRT methods, are used to construct the score scale. In any approach for constructing the score scale, the performance on the test or tests to be scaled is related to a single interim score scale. The interim score scale is transformed to a scale with specified properties. Within each general statistical approach, the specific procedures that are used depend on the design used for data collection. See Carlson (2011), Harris (2007), Kolen and Brennan (2004), Patz and Yao (2007), and Yen (2007) for more details on vertical scaling methods.

Scale-score equivalents by age group were developed for the subtests and process scores of the WAIS (Wechsler, 2008) by means of a vertical scaling process. The scaling test design was used in that all age groups of the normative sample responded to all of the same items that spanned the difficulty range of the WAIS. Statistical characteristics of the score distributions for each age group (mean, standard deviation, and skewness) were calculated and used to generate theoretical distributions for each of the reported normative age groups, yielding percentile ranks for each raw score. Two sets of scale scores are provided from the WAIS, one for the reference group (ages 20–34) and one based on the test taker's same age group. See Wechsler (2008) and Wilkins, Rolfhus, Weiss, and Zhu (2005) for more information on the method of inferential norming used to develop the age-based scale scores.

NORMS

In this section, norms and norm groups are defined. Then technical issues in the development of norms are considered, and illustrative examples of norming studies are described.

Norms and Norm Groups

Norms relate test scores to the performance of a group of test takers. National norms are based on drawing nationally representative samples of individuals at the age or educational level for which a test is designed. National norms are typically developed using a sampling plan that helps ensure that the sample accurately represents the population. National norms by age or grade are often provided for educational achievement and aptitude tests. National norms by gender may be provided for personality inventories and other noncognitive assessments. National norming studies are used to estimate test score characteristics, such as means, standard deviations, and percentile ranks, for a national population of test takers.

User norms are based on test takers who happen to take a test during a given time period or for a particular sample. These user norms cannot be viewed as nationally representative because they depend on who happens to take a particular test. User norms can facilitate score interpretation. For example, consider a student entering a college and completing the Learning and Study Strategies Inventory (Weinstein & Palmer, 2002), which is designed to measure students' awareness and use of learning and study strategies. The student's scores on the inventory can easily be compared with the national norms provided. However, also comparing the student's scores to the normative information for all students enrolled at the college (user norms) provides a closer

reference for interpreting and comparing the test taker's scores and will help to identify areas in which the college can best meet the test taker's needs.

Technical Issues in Development of National Norms

National norming studies are used to estimate test score characteristics, such as means, standard deviations, and percentile ranks, for a national population of test takers. The development of national norms involves drawing a representative sample of test takers from the national population. Sample survey methodology (Thompson, 2002) is used to design norming studies. In this section, some basic sampling concepts are considered. See Kolen (2006, pp. 180–183) for more details.

The *population of interest* is the population of test takers that the norms are intended to represent. The *population characteristics* or *population parameters,* such as means and percentile ranks for scores on a test, are the estimated quantities. A *sampling design* is the process that is used for sampling test takers from the population of interest. *Statistics* are the estimates of the population characteristics found from the sample.

Norming studies typically use a combination of sampling plans. In simple random sampling, each test taker in the population has an equal and independent probability of being included in the sample. In stratified random sampling, the population is divided into strata on the basis of test-taker characteristics, such as geographic region or public versus private school. A sample is drawn from each stratum. Statistics from each strata are often weighted differentially to estimate the population characteristic. Stratification reduces sampling error variance to the extent that the strata differ on the measured variable.

In systematic random sampling, every *n*th test taker is chosen from the population, after the first test taker is randomly chosen from among the first *n* test takers. If test takers are ordered randomly, then systematic random sampling is the same as simple random sampling. If the test takers are ordered on a variable related to the measured variable, then systematic random sampling can result in substantially lower sampling error than simple random sampling.

Cluster sampling involves sampling at the level of test-taker group. For example, schools might be sampled and then all students within a selected school tested. To the extent that the clusters differ, on average, on the test score of interest, cluster sampling requires testing more students than would be required with simple random sampling to achieve the same sampling error variance.

Most norming studies use a combination of sampling strategies. Simple random sampling is usually not practical for developing test norms. Specialists in sample survey design methodology also develop sampling designs and weights, as needed, so that the statistics that are calculated accurately estimate the population characteristics. In addition, indices of the precision of the estimates of the population characteristics are provided.

Illustrative Examples of National Norming Studies

Norming studies often use a combination of sampling strategies. In this section, procedures used in two national norming studies are described to illustrate the types and range of sampling procedures used in practice.

National norming studies are conducted for the NAEP, which is a national survey of educational achievement that provides information used by policymakers to inform decisions about education in the nation. NAEP is intended to broadly survey educational achievement in areas that include reading and mathematics. The breadth of these subject areas, and the desire to have adequate breadth of content surveyed, requires that each test taker take only a subset of the assessment. Scores are not reported to individual test takers. NAEP results are reported only at the group level, including the nation and various subgroups.

NAEP provides normative data on educational achievement at Grades 4, 8, and 12 in various subject matter areas (Institute of Education Sciences, 2011). In the development of norms, NAEP makes extensive use of sampling procedures described in technical manuals that accompany each assessment (e.g., Allen et al., 1999). Rust and Johnson (1992) described NAEP sampling that was used in 1986

through 1992. This discussion relies on their description as well as the summary provided by Kolen (2006, pp. 182–183).

A multistage sampling design was used with NAEP (Institute of Education Sciences, 2011). The first stage involved sampling *primary sampling units* (PSUs), which are geographical regions that contain a single metropolitan area, a single county that is not a metropolitan area, or a group of geographically contiguous counties. The United States was divided into approximately 1,000 PSUs. Because of their large size, some of the PSUs that contained a single metropolitan area were included in all NAEP samples (there were 34 of these PSUs in 1986–1992). The remaining PSUs were sorted into 60 strata by geographic region, whether the PSU was a metropolitan area, extent of minority population, and socioeconomic characteristics. One PSU was drawn randomly from each of these strata.

In the second stage, schools with students in the grade to be assessed were selected within the selected PSUs. The schools were chosen with probabilities proportional to the size of the schools, with the following exception. So that norms for important subgroups, such as African Americans and Hispanic Americans, were sufficiently precise, schools with high proportions of students from these subgroups were sampled at a higher probability. In the third stage, schools provided a list of students eligible for testing. Students were systematically sampled from these lists and assigned to test sessions.

Each student was administered only a subset of items from the entire pool of items. Within a test session, different students were administered different test questions. This procedure is generally referred to as *matrix sampling*. Different subsets of items were randomly assigned to students in each test session using a procedure referred to as *balanced incomplete block spiraling* (Beaton & Allen, 1992).

The records for students included in the sample were weighted to reflect the national population and to adjust for nonparticipation of students and schools. Weights were initially assigned as the reciprocal of the probability of selection of individual students for the assigned session. The weights were adjusted for the effects of nonparticipation of schools and students, and those weights that were extremely large were trimmed so that they did not overly influence the resulting norms. Poststratification was used to adjust the weights so that over the whole sample, they accurately reflected totals for population subgroups, as provided by the U.S. Census Bureau, defined by geographical region, race and ethnicity, and the relationship between student age and grade.

The weights were used to develop the norms on the NAEP for the U.S. population as a whole and for various subgroups. Associated estimates of precision were used to estimate the amount of sampling error present in the norms.

NAEP sampling was recently redesigned (Institute of Education Sciences, 2011). A sampling plan is now used within each state to produce state-level norms at a desired level of precision overall and for various subgroups within each state. The state samples are aggregated to produce the national sample.

Another well-documented norming study provides a contrast to the procedures used with NAEP. The WAIS norming sample (Wechsler, 2008) was recruited by market research firms located in eight cities across four major U.S. geographic regions (Northeast, Midwest, South, and West). Test-taker candidates were screened for issues that could affect cognitive test performance, and those with potentially confounding issues were excluded. Additionally, a representative proportion of individuals identified as intellectually gifted and intellectually disabled were included.

A stratified sampling plan was used to ensure that the normative sample included representative proportions of individuals according to several selected demographic variables—age, sex, race and ethnicity, self or parent education level, and geographic region.

The percentages of the resulting sample according to these demographic variables were compared with those provided by the U.S. Census Bureau from the October 2005 census. The comparison indicated that the resulting normative sample was nationally representative of the U.S. English-speaking population of individuals ages 16 to 90. No poststratification weighting was necessary. Subsequently, each chosen test taker was administered the entire set of WAIS items.

Both the NAEP and WAIS sampling procedures involved stratified random sampling, but for NAEP,

the first-stage sample was at the PSU level, whereas for the WAIS, the first-stage sample was at the level of larger geographic regions within the United States. A second difference is that a matrix sampling design was used to collect the exam data for the NAEP, whereas a complete sample design was used for the WAIS. The matrix sampling may have cut down on test takers' frustration and fatigue but required more complex procedures and logistics than administering all items to all test takers. These sorts of practical issues are always a concern when conducting national norm studies.

EQUATING

Alternate forms of tests are used for security purposes and so that individuals can take a test more than once. So that scores from alternate forms can be used interchangeably, such alternate forms are developed to be as similar as possible in content and statistical properties. Even when substantial effort is made to develop alternate forms that have scores that are equal in statistical characteristics, however, small differences are typically found. Equating methods are used to adjust for such differences. According to Kolen and Brennan (2004), "Equating is a statistical process that is used to adjust scores on test forms so that the scores on the forms can be used interchangeably" (p. 2). As this statement suggests, the goal of equating is to be able to use scores on alternate forms interchangeably.

To conduct adequate equating, alternate forms must be constructed to be very similar in content and statistical characteristics. Equating adjusts for small statistical differences in scores across alternate forms, not for differences in content or in the construct measured. Thus, detailed and well-articulated test development procedures that produce very similar alternate forms of tests are necessary for equating to be successful.

Equating has become an integral part of large-scale testing, especially for educational achievement and ability tests for which there is often a need for alternate test forms. A chapter by Flanagan (1951) provided the first extended treatment of equating, followed by a comprehensive chapter by Angoff (1971), and these references were updated by Petersen et al. (1989). Lord (1980) provided an overview of IRT equating. Holland and Rubin's (1982) edited book presented a number of research studies on equating methods. Kolen and Brennan (2004) provided a book-length introduction to equating. Von Davier, Holland, and Thayer (2004b) described a framework for traditional equating methods referred to as *kernel equating.* Recently, publications by Holland and Dorans (2006), Dorans et al. (2007), and von Davier (2011) considered equating in the more general context of methods for relating scores on assessments, which they referred to using the general term *linking,* that include equating as one type of linking. Livingston (2004) and Ryan and Brockmann (2009) provided overviews of equating designed to be helpful to practitioners. Although equating could in principle be used with alternate forms of measures of psychological constructs such as personality and attitudes and in medical assessment areas, few examples appear in the literature (see Dorans, 2007, and Orlando, Sherbourne, & Thissen, 2000, for two of the few such examples).

This section begins with a discussion of properties of equated scores. It continues with discussions of data collection designs, statistical equating methods, equating error, and practical issues in equating.

Equating Properties

Kolen and Brennan (2004, pp. 9–13) presented properties of equating that are described here. Similar lists of properties are presented in other references (e.g., Holland & Dorans, 2006; Petersen et al., 1989).

One property of equating is what Kolen and Brennan (2004) referred to as the *same-specifications property*. For this property to hold, the alternate test forms must be built to the same content and statistical specifications so that they are as similar as possible. This property is necessary for equating. Holland and Dorans (2006, p. 194) referred to the same construct property, which is a similar property.

The symmetry property (Lord, 1980) requires that the equating transformation be symmetric. For this property to hold, the transformation of scores from Form X to Form Y must be the inverse of the transformation of scores from Form Y to Form X. This property rules out regression as an equating

method because regression functions are not symmetric.

The equity property described by Lord (1980) is based on the consideration that it should be a matter of indifference to any test taker which form is administered. Lord statistically operationalized this property by stating that the distribution of observed equated scores for test takers of a particular true score should be the same on the alternate forms, and this property holds at all levels of true score. Lord went on to show that the equity property cannot hold unless the alternate forms are identical. Note that if equity holds, then alternate forms will be of equal reliability for a particular group of test takers. Dorans and Holland (2006, p. 19) included equal reliability of alternate forms as a property for equating. Brennan (2010) also discussed the need for similar reliability of alternate test forms as related to equity.

Morris (1982) suggested less restrictive versions of equity. He defined *first-order equity* by stating that the expected equated observed scores for test takers of a particular true score should be the same regardless of the alternate form taken and that this property holds at all true scores. This property implies that any test taker would be expected to earn the same equated score on alternate forms. Hanson (1991) provided practical conditions under which this property holds. Morris defined *second-order equity* by stating that the standard deviation of observed scores for test takers of a particular true score should be the same regardless of the alternate form taken and that this property holds at all true scores. This property implies that any test taker would be expected to be measured with the same precision on any alternate form.

Angoff (1971) described the observed-score equating property, which states that equated scores for alternate forms should have the same distribution in a population of test takers. When this property holds, for example, in a particular population, the same proportion of test takers would exceed a particular cut score on any alternate form.

For the group invariance property, the equating relationship is the same regardless of the group of test takers used in the equating. Under this property, for example, the equating relationship constructed using male test takers is the same as that constructed using female test takers. Research on the group invariance property (e.g., Angoff & Cowell, 1986; Harris & Kolen, 1986) suggested that this property holds for carefully constructed alternate forms. Dorans and Holland (2000) and Kolen and Brennan (2004) provided methodology for assessing group invariance.

The same specifications and symmetry properties are clear requirements for equating. The other properties should hold, at least approximately. Methods for assessing the other equating properties are discussed further in the Practical Issues of Equating section.

Designs

Designs for data collection are considered in this section, using terminology consistent with that of Kolen and Brennan (2004). In considering the equating designs, assume that scores on old Form Y have been transformed to scale scores. Scores on a new Form X are to be equated to scores on the previously equated Form Y and then to scale scores.

In the random-groups design, test takers are randomly assigned to take the forms. Such assignment is often done by packaging test booklets so that the first booklet is Form Y, the second Form X, and so forth. The booklets are then distributed in the order in which they are packaged. This spiraling process is intended to lead to random groups of test takers taking the forms. The groups taking each of the forms are considered to be randomly equivalent.

In the single-group with counterbalancing design, test takers are administered both forms to be equated. A random half of the test takers are administered Form Y followed by Form X, and the other half of the test takers are administered Form X followed by Form Y. Assuming that no differential order effects exist (i.e., taking Form X first vs. second has the same effect on scores as taking Form Y first vs. second), the equating functions for the two orders are averaged to produce the equating function that is used. Sometimes the single-group design is used without counterbalancing, although doing so requires an assumption that order effects are not consequential.

The random-groups design is preferable to the single-group design whenever equating is to be done

using operational test takers because the test takers only have to take one form. In situations in which two forms can be given and no differential order effects exist, the single-group design with counterbalancing can often lead to greater equating precision than the random-groups design because each test taker serves as his or her own control.

Sometimes developers of assessments need to conduct equating as part of an operational assessment program. It might be impossible, in this case, to administer two forms to the same test takers or to administer two forms to random groups of test takers. In this case, a common-item nonequivalent-groups design (referred to as a *nonequivalent-groups-with-anchor test design* by Holland & Dorans, 2006) can be used. In this design, the groups of test takers taking the new Form X, referred to as Group 1, is considered to be different, that is nonequivalent, to the group that took the old Form Y, referred to as Group 2. A set of common items, V, is administered to both groups. In the external common-item version of this design, the score on V does not contribute to scores on Form X or Form Y. When using external common items, the common items are often administered in a separately timed section. In the internal common-item version of this design, the common items contribute to scores on both Form X and Form Y. When using the internal common-item version, the common items are typically interspersed throughout the forms.

In the common-item nonequivalent-groups design, scores on the common items provide a clear indication of how the test takers in the group who took Form X differ on the assessed construct from the test takers in Group 2 who took Form Y. Strong statistical assumptions are made so that the information about group differences in scores on the common items can be used to estimate statistical differences in scores on the complete forms. For the scores on the common items to adequately reflect group differences on the complete form, it is important that the content of the common items proportionally represent the content of the complete forms and that the common items be presented in the same context on both forms. For example, the common items should be in nearly the same position on both forms, and the wording and formatting of the items should be the same. Even though the requirements for using the common-item nonequivalent-groups design are quite stringent from an assessment development perspective, this design is often used because there are many situations in which the random-groups and single-groups designs with counterbalancing cannot be used.

Both traditional and IRT statistical methods can be used with the three designs just described, as discussed later in this chapter. Additional designs can be used with IRT. For example, with IRT, calibrated item pools are often developed in which a large number of items have IRT item parameter estimates on a common scale.

In the common-item-equating-to-an-IRT-calibrated-item-pool design, a new Form X is constructed that contains a set of items that are already in the calibrated item pool as well as some new items. After administration of Form X, test-taker performance on the common items and the IRT model assumptions are used to assess the differences between the group of test takers taking the new form and the group used to establish the IRT scale for the item pool. The statistical assumptions of IRT are used to convert scores on Form X to scores on base Form Y and then to scale scores. On the basis of the same reasoning used with the common-item nonequivalent-groups design, the common items should proportionally represent the content of Form X and the items should be placed in an item position similar to that used when they were administered previously.

In the item-preequating-to-an-IRT-calibrated-item-pool design, new Form X is constructed from items that are in the IRT-calibrated item pool. Because item calibrations for these items exist in the pool, the conversion of Form X scores to scores on base Form Y and then to scale scores can be estimated from the item parameter estimates that already exist in the pool. This design has the practical benefit that score conversions can be developed before the form is ever administered and scores can be provided immediately after the form is taken. With this design, it is important that Form X be built to the same content and statistical specifications as previous forms, which leads to

equating that tends to be robust to violations of the IRT assumptions that often occur in practice (Kolen & Brennan, 2004, pp. 205–207). It is also important that the items be placed in an item position similar to that used when previously administered. One potential drawback of this design is that problems that are found with any items (e.g., with content or with printing) after the administration cannot be easily corrected because scores are often reported to test takers immediately after they take the form.

Statistical Equating Methods

In this section, traditional linear and equipercentile equating methods are introduced in the relatively simple random-groups design followed by a discussion of these methods with the common-item nonequivalent-groups design. For presentation of the traditional methods, it is assumed that the score for each item is discrete and that the total score to be equated is based on a sum of the item scores, which are referred to as *summed scores*. IRT methods for both designs are then presented. The section concludes with a discussion of the use of IRT in the two designs that make use of IRT-calibrated item pools. Note that all of the methods described in this section are presented in much greater detail in Kolen and Brennan (2004).

Linear methods for the random-groups design. In linear methods of equating, a linear transformation is found that transforms scores on Form X to scores on Form Y so that the scores have the same mean and standard deviation for a particular group of test takers. The form of the conversion equation is a straight line.

Define X as summed score on Form X, Y as summed score on Form Y, x as a particular summed score on Form X, y as a particular summed score on Form Y, $\mu(X)$ as the mean summed score on Form X, $\mu(Y)$ as the mean summed score on Form Y, $\sigma(X)$ as the standard deviation of summed scores on Form X, and $\sigma(Y)$ as the standard deviation of summed scores on Form Y. The linear method is developed by setting standardized scores (z scores) on the two forms equal. Defining $l_Y(x)$ as the linear equating transformation that transforms scores on Form X to scores on Form Y, Kolen and Brennan (2004, p. 31) showed that

$$l_Y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right]. \quad (11.1)$$

This equation is expressed as a linear equation in slope and intercept form, with the slope being the ratio of the standard deviations and the intercept the term in square brackets. Note that the linear equating function depends only on the means and standard deviations of the scores on Form X and Form Y. Kolen and Brennan (2004, pp. 32–33) demonstrated that when the scores on Form X are transformed to scores on Form Y using this transformation, the transformed scores have the same mean and standard deviation as the original Form Y scores. They also demonstrated that this equation is symmetric. That is, to find the Form X equivalents of Form Y scores, this equation can be solved for x.

The linear equating equation can lead to transformed scores that are outside of the range of possible scores on Form Y. In this case, the transformed scores are for practical purposes often truncated to be in the appropriate range. Note that the linear equating equation is expressed in terms of parameters. In practice, sample means and standard deviations are used in place of the parameters.

Equipercentile methods for the random-groups design. In equipercentile methods of equating, a curvilinear transformation, $e_Y(x)$, is found that transforms scores on Form X to scores on Form Y so that the scores have approximately the same distribution for a particular group of test takers. Define F as the cumulative distribution function for X and G as the cumulative distribution function for Y. For example, $F(x)$ represents the proportion of scores below the score x. Braun and Holland (1982) indicated that when X and Y are continuous random variables, the following function can be used to transform scores on Form X to scores on Form Y so that the transformed scores have the same distribution as the scores on Form Y:

$$e_Y(x) = y = G^{-1}\left[F(x)\right], \quad (11.2)$$

where G^{-1} is the inverse of the cumulative distribution function G. That is, G^{-1} is the score y that has a particular proportion of scores below. Kolen and

Brennan (2004, p. 37) demonstrated that this equation is symmetric.

In most equating applications, the scores are discrete integer scores, such as when summed scores are used. With such discrete scores, G^{-1} often cannot be found. To deal with this issue, percentiles are typically used in place of G^{-1} and percentile ranks are typically used in place of $F(x)$. With percentiles and percentile ranks, discrete scores are treated as if they are uniformly distributed on an interval from 0.5 score point below each integer score to 0.5 score point above (see Kolen & Brennan, 2004, pp. 43–46).

To implement equipercentile equating, for each score x the percentile rank of the score is found, which is the percentage of scores below score x. Then, the score on Form Y that has that percentage of scores below it is the equipercentile equivalent of score x. Kolen and Brennan (2004) demonstrated that the distribution of Form X scores equated to the Form Y scale is approximately equal to the distribution of Form Y scores. In addition, they demonstrated that the equipercentile equivalents that result are within the range of scores on Form Y plus or minus 0.5. In practice, estimates of the percentiles and percentile ranks are used in place of the parameters.

Because so many parameters must be estimated in equipercentile equating (percentiles and percentile ranks at each score point), the estimates of the equipercentile equivalents are subject to considerable sampling error. In addition, the equipercentile equating function often appears irregular. For this reason, smoothing methods are often used with equipercentile equating. In postsmoothing, the equipercentile equivalents are smoothed. Kolen and Brennan (2004, pp. 84–91) described and illustrated a postsmoothing method developed by Kolen (1984). In presmoothing, the score distributions are smoothed and then the smoothed distributions are equated using equipercentile methods. Von Davier, Holland, and Thayer (2004b), Holland and Thayer (2000), and Kolen and Brennan (2004) illustrated a presmoothing method that uses log-linear models.

Kolen and Brennan (2004, p. 72) indicated that smoothing methods should be accurate and flexible, have a statistical framework for evaluating fit, and have an empirical research base that demonstrates that it improves estimation of equating relationships. The postsmoothing method they described and the log-linear presmoothing method meet these criteria.

Von Davier et al. (2004b) described a kernel method that can be used in place of percentiles and percentile ranks. At each score point, the discrete scores are treated as if they are normally distributed. The score distributions that result from application of the kernel method are continuous and equipercentile equating can be accomplished directly. Von Davier et al. paired the kernel method with log-linear smoothing to produce what they referred to as *kernel equating*. Because kernel equating uses the normal distribution to deal with the discrete scores, the range of score equivalents is infinite.

The equipercentile method tends to produce equivalents that are similar to the linear method in the middle of the score distribution. However, the methods can produce very different equivalents, especially at more extreme scores, when the shapes of the score distributions for the forms to be equated differ considerably.

Linear methods for the common-item nonequivalent-groups design. As previously indicated, when equating is conducted using the common-item nonequivalent-groups design, the new Form X is administered to test takers from Group 1, old Form Y is administered to test takers from Group 2, and a set of common items, V, is administered to both groups. When the common items contribute to the test taker's score on the test form, they are referred to as *internal common items*. When the common items do not contribute to the test taker's score on the common items, they are referred to as *external common items*.

Test-taker scores on the common items provide direct evidence of the difference on the construct being assessed between Group 1 and Group 2. For example, a higher mean on the common items for Group 1 than for Group 2 is direct evidence that Group 1 is, on average, higher on the construct being assessed than Group 2. In equating using the common-item nonequivalent-groups design, the differences between the two groups that are observed, along with statistical assumptions about

the relationships between scores on the common items and scores on the test forms, are used to estimate the differences in scores on the two forms. Various equating methods for this design exist, and they differ in the statistical assumptions that are made. Note that with this design, Form Y scores are not observed in Group 1 and Form X scores are not observed in Group 2.

In the Tucker method, which was attributed to Ledyard Tucker by Gulliksen (1950), linear equating is conceived of for a synthetic group of test takers that is a weighted combination of Groups 1 and 2. In this case, the linear equating function is the same as that in Equation 11.1, but with a synthetic group subscript added to each term. To find the parameters for the synthetic group, an assumption is made that the slope and intercept of the unobserved linear regression of *X* on *V* in Group 2 is equal to the linear regression of *X* on *V* in Group 1. In addition, it is assumed that the unobserved linear regression of *Y* on *V* in Group 1 is equal to the linear regression of *Y* on *V* in Group 2.

The Levine equally reliable method was introduced by Levine (1955) and assumes that true scores of *X* and *V* and true scores of *Y* and *V* are perfectly correlated in Groups 1 and 2, implying that scores on the total tests and the common items are measuring exactly the same construct. This method also assumes that the slope and intercept of the unobserved linear regression of true scores for *X* on true scores for *V* in Group 2 is equal to the linear regression of true scores for *X* on true scores for *V* in Group 1. In addition, it is assumed that the unobserved linear regression of true scores for *Y* on true scores for *V* in Group 1 is equal to the linear regression of *Y* on *V* in Group 2. Derivations for both the Tucker method and the Levine equally reliable method are provided by Kolen and Brennan (2004), along with the resulting equations (p. 122).

Kolen and Brennan (2004, p. 147) mentioned a chained linear method, which was described in more detail by Holland and Dorans (2006, p. 208). In this method, scores on *X* are linked to scores on *V* for Group 1 using a linear equating function such as that defined in Equation 11.1 and symbolized by $l_{v1}(x)$. Scores on *V* are linked to scores on *Y* for Group 2 using a linear equating function such as that defined in Equation 11.1 and symbolized by $l_{Y2}(v)$, and these two functions are chained together to produce the chained linear equating function

$$l_{Ychain}(x) = l_{Y2}\left[l_{v1}(x)\right]. \quad (11.3)$$

Note that unlike the Tucker (Gulliksen, 1950) and Levine (1995) observed score methods, the chained method does not use the concept of the synthetic group. In addition to the method described here, Kolen and Brennan and Holland and Dorans discussed a Levine true score equating procedure that was introduced by Levine (1955).

Note that the Tucker method (Gulliksen, 1950) involves assumptions that linear regressions are the same in Groups 1 and 2. Given the typical population dependence of regression functions, the assumptions for this method might break down when Groups 1 and 2 differ considerably on the construct assessed by the two forms. The Levine method makes an assumption about true scores on the common items being perfectly correlated with true scores on the total test forms. This assumption for the Levine method might break down whenever the content of the common items is not representative of the content of the test forms. Von Davier et al. (2004a) indicated that the assumption for the chained method is that the linear linking functions that contribute to Equation 11.3 are the same across populations. Each of these linear methods typically leads to different linear equating functions because of the different assumptions that are made. In practice, it is advisable to conduct equating using each of these methods and to evaluate the suitability of the assumptions and the results for the particular equating that is conducted.

Equipercentile methods for the common-item nonequivalent-groups design. As with the Tucker method (Gulliksen, 1950), the equipercentile equating relationship is defined for the synthetic group of test takers for the frequency estimation equipercentile method. In the frequency estimation equipercentile method, an assumption is made that the unobserved conditional distribution of scores of *X* on *V* for Group 2 is equal to the observed conditional distribution of *X* on *V* for Group 1. In addition, it is assumed that the unobserved conditional

distribution of scores of *Y* on *V* for Group 1 is equal to the conditional distribution of observed scores of *Y* on *V* for Group 2. By making these assumptions, it is possible to express the marginal distributions of *X* and *Y* for the synthetic group. These distributions are equated using Equation 11.2. Kolen and Brennan (2004, pp. 135–139) provided equations that can be used to implement this method.

The chained equipercentile method (Holland & Dorans, 2006, p. 208; Kolen & Brennan, 2004, pp. 145–147) works much as does the chained linear methods. For Group 1, *X* is linked to *V* using equipercentile method. For Group 2, *V* is linked to *Y* using equipercentile methods. These two linking functions are chained as was done in chained linear methods to produce the chained equipercentile equating function.

Observed frequency distributions are used to implement these methods in practice. Both presmoothing and postsmoothing methods can be used with frequency estimation and chained equipercentile equating methods, as discussed by Kolen and Brennan (2004, pp. 142–143).

Item response theory equating methods. IRT provides a psychometric foundation for many testing applications (de Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Lord, 1980; van der Linden & Hambleton, 1997; see also Chapter 6, this volume). IRT makes strong psychometric assumptions that allow for the solution of many practical problems in measurement. For unidimensional IRT models, a unidimensionality assumption is made that implies that all items measure the same construct that is symbolized by θ. Such models also assume local independence, meaning that conditional on θ, test takers' item responses are statistically independent.

IRT models responses at the item level. Let U_i represent the random variable score on item *i* and u_i a particular score on that item. A central concept of IRT is the category response function that relates the probability of earning a particular score on an item to θ and is symbolized by $P(U_i = u_i \mid \theta)$. For dichotomously scored items on ability or achievement tests, items are typically scored 0 for an incorrect response and 1 for a correct response, and the item response function is typically defined as $P(U_i = 1 \mid \theta)$, which is the probability of correctly answering the item. For such items, the probability of correctly answering an item is sometimes modeled using the three-parameter logistic model that has difficulty, discrimination, and lower asymptote parameters (Lord, 1980) for each item (for more details on IRT approaches, see Chapter 6, this volume). Special cases of this model include the two-parameter logistic model, which assumes that the pseudo–chance-level parameter is 0, and the Rasch model, which assumes that the pseudo–change-level parameter is zero and that the discrimination parameter is 1. Various models also exist for polytomously scored items including the graded response model (Samejima, 1997), the generalized partial credit model (Muraki, 1997), and the partial credit Rasch model (Masters & Wright, 1997). Whenever IRT is used in equating, it is also necessary to assume that the same theta is measured by Form X and Form Y and that the same theta is measured in Groups 1 and 2.

Item response theory scale linking. The location and spread of the theta scale is arbitrary, and with many computer programs used with these models, the scale is set so that theta has a mean of 0 and a standard deviation of 1 for the group of examinees. Sometimes the item parameters are estimated for Form X and Form Y in separate computer runs. In this case, when the random-groups or single-group with counterbalancing design is used, the estimated parameters are on the same scale that has a mean of 0 and a standard deviation of 1.

When the common-item nonequivalent-groups design is used and the item parameters for Form X and Form Y are estimated separately, a linear scale transformation is needed to place the estimates of theta and the item parameter estimates on the same scale using a linear transformation (Kolen & Brennan, 2004, p. 162). Procedures for estimating these transformation constants are described in detail by Kolen and Brennan (2004, pp. 173–175, 215–218).

When the common-item nonequivalent-groups design is used and the parameters for Form X and Form Y are estimated concurrently using a computer program such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) or PARSCALE

(Muraki & Bock, 2003), the scale is set so that the mean and standard deviation of theta for one of the groups, say Group 2, can be set to 0 and 1, respectively, and the mean and standard deviation of theta for the other group of test takers is estimated by the computer program. In this case, the concurrent calibration process results in parameter estimates for the common-item nonequivalent-groups design being on the same scale, and no further scale transformation is needed.

For the common-item nonequivalent-groups design, concurrent calibration has the benefit of avoiding the scale transformation step. In addition, concurrent calibration makes use of all of the information on both groups of test takers and both forms in estimating item parameters, which can lead to more stable parameter estimation. However, Kolen and Brennan (2004, pp. 173–175) pointed out that there can be convergence problems with concurrent calibration, and there is some evidence that concurrent calibration is less robust to violations of the IRT assumptions than separate calibration. In addition, in many testing programs, item parameter estimates exist for the previously administered Form Y so there is no need to estimate them again using concurrent calibration.

Item response theory true and observed score methods. When the IRT theta scale is the scale used for score reporting, no further equating transformation is needed. However, if there is a desire to equate summed scores on one form to summed scores on another form, then IRT true- and observed-score equating methods can be used after all of the item parameters and theta estimates are placed on the same scale using the scale linking procedures just mentioned.

In IRT true-score equating, true scores on the two forms are related to one another. For tests that consist of dichotomously scored items, the true scores on Form X and Form Y, test response functions are found by summing the item response functions over items on a form to produce a true score on that form. This sum represents the expected number of items that an examinee of a particular theta would be expected to correctly answer on each of the forms. A true score on Form X is considered to be equated to a true score on Form Y when both true scores are associated with the same theta. This process is described by Kolen and Brennan (2004, pp. 176–178, 219–220).

The IRT observed-score equating method uses the IRT model to produce smooth score distributions for Form X and Form Y for a population of test takers. These smoothed distributions are then equated using equipercentile methods. Kolen and Brennan (2004, pp. 181–184, 194–198) provided a detailed description of IRT observed-score equating methods.

IRT true- and observed-score equating methods have the benefit of providing methods that parallel the traditional methods, so they can be used when both IRT and traditional methods are used. IRT true-score methods have the advantages of less computational burden and of not depending on the distribution of theta, compared with IRT observed-score methods. However, there is no justification for applying the IRT true-score equating relationship to observed scores. Kolen and Brennan (2004, pp. 184–185) reviewed research, including studies by Lord and Wingersky (1984) and Han, Kolen, and Pohlmann (1997), that has suggested that the two methods produce similar results along most of the score scale.

Item response theory scale-linking and equating for an item response theory–calibrated item pool. Whereas the traditional equating methods require equating scores from a new form to an old form, IRT methods are more flexible because they allow equating of scores from a new form through an item pool rather than to scores on a particular old form. To use the common-item-equating-to-an-IRT-calibrated-item-pool design that was described earlier in this chapter, items are included on new Form X that were previously administered and that have item parameter estimates on the theta scale in the item pool. The Form X items are calibrated using an IRT method. Then the scale-linking methods described earlier in this chapter are applied using parameter estimates for the items that are in common with the pool. The item parameter estimates and estimates of theta or distributions of theta are then transformed to the theta scale for the item

pool. If desired, IRT true- or observed-score methods can be used to equate summed scores on Form X to summed scores on a base form.

For the item preequating with an IRT-calibrated item-pool design, the new Form X is constructed using items that are already calibrated on the basis of previous use. Because the items are already calibrated, conversions to scale scores can be known before Form X is administered intact. To expand the pool over time and to allow for items to be retired from the pool after a certain number of uses, new items that do not contribute to the score on Form X are often administered when this design is used. These new items are calibrated after the administration, with their parameter estimates linked to the theta scale for the item pool using the scale-linking methods described earlier in this chapter.

Equating Error

Two general sources of equating error occur when equating is conducted. *Systematic equating error* occurs whenever statistical assumptions do not hold; *random equating error* is error resulting from sampling test takers from a population of test takers. Random error becomes smaller as sample size increases.

Systematic error is best controlled through the design of the equating data collection. For example, when the random-groups design is used, systematic error can be controlled by ensuring that the procedures for randomly assigning individuals to test forms are adequately implemented. As another example, with the common-item nonequivalent-groups design, it is assumed that the common items behave similarly on the old and new forms. Placing common items in the same position on the old and new forms is one way to make sure that the common items are behaving similarly on the two forms. In addition, all of the statistical methods used with the common-item nonequivalent-groups design require that strong statistical assumptions be made. These assumptions likely hold better, and hence minimize systematic error, under the following conditions: when (a) the forms to be equated are as similar as possible, (b) the common items proportionally represent the complete forms, and (c) the groups taking the old forms are fairly similar to one another. Whenever these conditions do not hold, substantial systematic error might be present. Thus, systematic error is best controlled through design of equating studies to minimize the possibility of systematic error.

Random equating error is typically indexed by the standard error of equating. Standard errors of equating depend on the data collection design, the equating methods, the population of test takers, the sample size, and the score level of interest. Conceptually, the standard error of equating is the standard deviation of equated scores over repeated samples of test takers from the population or populations of test takers. Formulas exist for the standard error of equating for the commonly used traditional and IRT methods and designs (Kolen & Brennan, 2004, pp. 245–253). Von Davier et al. (2004b) provided a comprehensive procedure for estimating standard errors of equating for kernel equating methods. These methods can be used to estimate and document the amount of random error in equating, and they can be used to estimate the sample sizes needed to achieve a desired degree of equating error. In general, standard errors of equating decrease as sample size increases.

Kolen and Brennan (2004, pp. 235–245) described how to use the bootstrap resampling method (Efron & Tibshirani, 1993) to estimate bootstrap standard errors of equating. These methods can be used with any of the equating designs and methods described in this chapter. They can also be used to estimate standard errors of equating for rounded scale scores.

Practical Issues in Equating

In this section, some important practical issues in equating are reviewed. These practical issues and others are also reviewed by Cook (2007); Dorans, Moses, and Eignor (2011); Holland and Dorans (2006); Kolen and Brennan (2004); Petersen (2007); Petersen et al. (1989); and von Davier (2007).

Equating and test development. According to Mislevy (1992),

> test construction and equating are inseparable. When they are applied in concert, equated scores from parallel test forms

> provide virtually exchangeable evidence about students' behavior on the same general domain of tasks, under the same standardized conditions. When equating works, it is because of the way the tests are constructed. (p. 37)

Thus, the construction of test forms that are as similar as possible is a necessary condition for adequate equating. In educational achievement tests, comparability of forms is accomplished by using detailed content and statistical specifications that lead to alternate forms that are very similar in content and difficulty. Having detailed test specifications is important for equating forms for other psychological constructs.

Common items. When using the common-item nonequivalent-groups design or the common-item-equating-to-an-IRT-calibrated-item-pool design, it is important that the common items represent the complete test forms in content and statistical characteristics so that scores on the common items adequately represent group differences. In addition, it is important to design the forms so that the common items behave the same way when administered in any of the alternate forms. Having common items presented in the same position on the test and in exactly the same format help to make sure that they will behave the same way in the alternate forms. When the common-item nonequivalent-groups design is used, statistical checks can be used to identify common items that are behaving differently in the old and new forms (Kolen & Brennan, 2004, p. 271).

Equating properties. Equating properties were discussed earlier in this chapter. These properties can be checked. Kolen and Brennan (2004, pp. 301–306) showed how to assess the extent to which first- and second-order equity holds using an IRT model as a psychometric foundation. Van der Linden (2000, 2006) provided an approach for considering the equity property. Tong and Kolen (2005) found that the equity properties typically held well for the equatings they investigated, except in those cases in which the test forms differed considerably in difficulty. The group invariance property can be assessed using procedures presented by Dorans and Holland (2000), Dorans (2004b), and Kolen and Brennan (2004, pp. 437–465). Dorans (2004a) edited a special journal issue that illustrated that population invariance tended to not hold as tests whose scores were linked differed in the construct that was being assessed. However, when equating test forms that are very similar to one another, the equating is typically found to not depend very much on the population. See Dorans (2004b) for more detail.

Choosing among equating results. When conducting equating in practice, it is often necessary to choose among results from different equating methods. Such choices often need to be made without strong evidence as to which result is preferable. One question to address is the following: Which method has statistical assumptions that are most likely to be met for this particular equating? For example, if the common-item nonequivalent-groups design is being used and the group differences are large, then methods based on regression assumptions such as the Tucker method might be less robust than chained methods. Another question is how the results from this equating compare with historical information. For example, if pass rates for a particular administration were around 75% and one method leads to a pass rate of 60% and another to a pass rate of 74%, then the second method might be preferred. In practice, the entire equating context needs to be considered, and it can be difficult to have substantial confidence in choice of method.

Quality control. Quality control procedures are crucial when conducting equating and are often the most time-consuming part of operational equating in large-scale testing programs (Kolen & Brennan, 2004, p. 309). Quality control procedures include checking that administration conditions are followed properly, answer keys are properly specified, items appear as intended, equating procedures are followed correctly, score distributions and statistics are consistent with those observed in the past, and the correct conversion table or equations are used.

Equating and constructed-response tasks. In educational testing, the use of constructed-response tasks that are scored by human raters, such as questions that require extended written responses, often serve as one or more components on a test that is

also composed of objective test items. As Kolen and Brennan (2004, pp. 320–323) pointed out in their review of research in this area, the extensive use of constructed-response tasks creates complications for equating methods. Because of the memorability of these tasks, it is often not possible to repeat such tasks on alternate forms. In such cases, it may be necessary to use only multiple-choice questions as common items, even though the common items are then not representative of the total test. This situation causes challenges for equating, and it is an area that is the focus of current applied research (see, e.g., Kim, Walker, & McHale, 2010).

Mode of administration. There has also been a trend away from the administration of tests on paper to administration on computer. The question then arises about whether scores from one mode of administration can be used interchangeably with those from another mode. Kolen and Brennan (2004, p. 317) reviewed studies of mode effects and found mixed results. Some studies found noticeable effects, and others did not. Overall, it appears that mode effects are complex and likely depend on various aspects of the testing situation, including the type of construct being assessed and the computer interface used. Eignor (2007) reviewed a number of studies of mode effects and discussed practical issues associated with mode effects, with a focus on the design of studies for linking scores on tests administered in different modes.

Methods for Linking Scores on Different Assessments

The term *linking* is used as a general term to refer to relating scores on different assessments (Dorans et al., 2007; Holland, 2007; Holland & Dorans, 2006). Linking encompasses the equating, vertical scaling, and relating scores across modes of administration processes that have already been discussed in this chapter. Linking also encompasses the process of scaling tests from a battery to have similar statistical properties, as was discussed earlier.

Sometimes linking scores on different tests is desirable. Situations considered by Dorans et al. (2007) included linking scores from one edition of a test to another after the test specifications have been changed and linking scores on tests intended for groups, such as the NAEP, to tests intended for individuals, such as an achievement test. As another example, Dorans, Lyu, Pommerich, and Houston (1997) described linking scores on the SAT and ACT that are used for college admissions in the United States. Many colleges accept either SAT or ACT scores and have a need to have comparable scores for admissions purposes for them. When scores on two assessments to be linked are highly related, equipercentile procedures are often used to link the scores, and the linking process is referred to as a *concordance*. When scores on two assessments to be linked are not highly related, Holland and Dorans (2006) indicated that prediction methods should be used to link the scores on the tests.

CONCLUSIONS

The variety of score scales that are used with psychological tests were described and illustrated in this chapter. As emphasized, the primary purpose of the use of score scales is to facilitate score interpretation by test users through the incorporation of normative and content information.

As discussed, equating procedures are used in many and various testing programs because of a need for alternate test forms for security purposes and so that individuals can be tested more than one time. The goal of any equating method is to be able to use scores on alternate forms interchangeably. Equating has the strong requirement that the alternate forms be developed to the same content and statistical specifications. Equating requires a design for data collection and the use of statistical procedures. A variety of data collection designs and both traditional and IRT statistical procedures were described in this chapter.

For many testing programs, a score scale is constructed and normative or content information is incorporated when the testing program is initiated. As new test forms are developed, the scores on the new forms are equated, and the resulting scale scores are considered interchangeable, regardless of the test form administered. This scaling and equating process is a key component of many psychological tests and assessments.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report.* Washington, DC: National Center for Education Statistics.

American College Testing. (2007). *ACT Assessment technical manual.* Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23,* 327–345. doi:10.1111/j.1745-3984.1986.tb00253.x

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*, 191–204. doi:10.2307/1165169

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York, NY: Academic Press.

Brennan, R. L. (2010). *First-order and second-order equity in equating* (CASMA Research Report No. 30). Iowa City: University of Iowa.

Carlson, J. E. (2011). Statistical models for vertical linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 59–70). New York, NY: Springer.

Cook, L. L. (2007). Practical problems in equating test scores: A practitioner's perspective. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 73–88). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_5

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction.* Englewood Cliffs, NJ: Prentice-Hall.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Dorans, N. J. (Ed.). (2004a). Assessing the population sensitivity of equating functions [Special issue]. *Journal of Educational Measurement, 41*(1).

Dorans, N. J. (2004b). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41,* 43–68. doi:10.1111/j.1745-3984.2004.tb01158.x

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research, 16*, 85–94. doi:10.1007/s11136-006-9155-3

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281–306. doi:10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University, 73*(2), 24–32.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). New York, NY: Springer.

Dorans, N. J., Pommerich, M., & Holland, P. (Eds.). (2007). *Linking and aligning scores and scales.* New York, NY: Springer. doi:10.1007/978-0-387-49771-6

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement*, *22,* 15–25.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability 57). New York, NY: Chapman & Hall.

Eignor, D. R. (2007). Linking scores derived under different modes of administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135–159). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_8

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley. doi:10.1037/13240-000

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150. doi:10.2307/2086306

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10*, 105–121. doi:10.1207/s15324818ame1002_1

Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design.

Journal of Educational Statistics, 16, 93–100. doi:10.2307/1165113

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–251). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_13

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*, 35–43. doi:10.1177/014662168601000103

Hathaway, S. R., & McKinley, J. C. (1989). *Minnesota Multiphasic Personality Inventory—2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_2

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education.

Holland, P. W., & Rubin, D. B. (1982). *Test equating.* New York, NY: Academic Press.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa tests: Guide to research and development*. Itasca, IL: Riverside.

Hsu, L. M. (1984). MMPI T scores: Linear versus normalized. *Journal of Consulting and Clinical Psychology, 52,* 821–823. doi:10.1037/0022-006X.52.5.821

Institute of Education Sciences. (2011). *National Assessment of Educational Progress*. Retrieved from http://nces.ed.gov/nationsreportcard/about

Kane, M. (2008). The benefits and limitations of formality. *Measurement: Interdisciplinary Research and Perspectives, 6,* 101–108. doi:10.1080/15366360802035562

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47*, 36–53. doi:10.1111/j.1745-3984.2009.00098.x

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*, 25–44. doi:10.2307/1164830

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: American Council on Education.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin RB-55-23). Princeton, NJ: Educational Testing Service.

Livingston, S. A. (2004). *Equating test scores (without IRT*). Princeton, NJ: Educational Testing Service.

Lohman, D. F., & Hagen, E. P. (2002). *Cognitive Abilities Test (Form 6): Research handbook*. Itasca, IL: Riverside.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461. doi:10.1177/014662168400800409

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York, NY: Springer.

McCrae, R. R., & Costa, P. T., Jr. (2010). *Professional manual for the NEO Inventories.* Odessa, FL: Psychological Assessment Resources.

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives, 6,* 7–24. doi:10.1080/15366360802035489

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 169–191). New York, NY: Academic Press.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York, NY: Springer-Verlag.

Muraki, E., & Bock, R. D. (2003). *PARSCALE* (Version 4.1). Mooresville, IN: Scientific Software.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment, 12,* 354–359. doi:10.1037/1040-3590.12.3.354

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253–272). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_14

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M.

Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_4

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: Macmillan.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rust, K. F., & Johnson, E. G. (1992). Sampling and weighting in the National Assessment. *Journal of Educational Statistics, 17*, 111–129. doi:10.2307/1165165

Ryan, J., & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on classical test theory and item response theory*. Washington, DC: Council of Chief State School Officers.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer-Verlag.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York, NY: Wiley.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 1–76). New York, NY: Wiley.

Thompson, S. K. (2002). *Sampling.* New York, NY: Wiley.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433–451. doi:10.1037/h0073357

Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychological Review, 35*, 175–197. doi:10.1037/h0072902

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29,* 418–432. doi:10.1177/0146621606280071

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika, 65,* 437–456. doi:10.1007/BF02296337

van der Linden, W. J. (2006). Equating error in observed-score equating. *Applied Psychological Measurement, 30,* 355–378. doi:10.1177/0146621606289948

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

von Davier, A. A. (2007). Potential solutions to practical equating issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 89–106). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_6

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking.* New York, NY: Springer. doi:10.1007/978-0-387-98138-3

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41,* 15–32. doi:10.1111/j.1745-3984.2004.tb01156.x

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). *The kernel method of test equating.* New York, NY: Springer.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth edition: Technical and interpretive manual.* San Antonio, TX: Pearson.

Weinstein, C. E., & Palmer, D. R. (2002). *User's manual for the Learning and Study Strategies Inventory* (2nd ed.). Clearwater, FL: H & H.

Wilkins, C., Rolfus, E., Weiss, L., & Zhu, J. J. (2005, April). *A new method for calibrating translated tests with small sample sizes*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116. doi:10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: Mesa Press.

Yen, W. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). New York, NY: Springer. doi:10.1007/978-0-387-49771-6_15

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299–325. doi:10.1111/j.1745-3984.1986.tb00252.x

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG* (Version 3.0). Morrisville, IN: Scientific Software.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20,* 15–25. doi:10.1111/j.1745-3992.2001.tb00059.x

CHAPTER 12

BASIC ISSUES IN THE MEASUREMENT OF CHANGE

John J. McArdle and John J. Prindle

There should be no doubt that psychology's most important decisions are appropriately made on the basis of the answers to the question, "Will it change things"? Of course, many key decisions are made by key people, and these decisions have certainly had an impact on entire financial and educational systems. Unfortunately, these kinds of decisions are most often made without an empirical or objective basis, either because there are no data available or because the available data have not been used. This makes it very easy to be critical of the typical subjective approach to decision making.

The ball is back in psychology's court when researchers are given the rare opportunity to evaluate change on an empirical—seemingly objective—basis. Researchers typically start by requesting at least two occasions of the same data—that is, repeated measures. They then apply the repeated-measures analysis of variance (RANOVA) to these data and announce the findings. Unfortunately, if they can actually obtain such difficult-to-obtain data, they often become the ones who are criticized, often by their own peers. The first problem raised is that the data at Time 1 are often not exactly repeated at Time 2; the conditions of measurement have changed, often substantially; or the participants simply do not want to do it again a second time. This means that often researchers have collected very hard-to-collect data and that just as often they have been left scrambling for exactly what they should do. These researchers initially thought they were in the possession of very well-known and highly robust procedures for the so-called analysis of change, based on the standard RANOVA model, but soon found out that many of the data analysis techniques used have inherent problems as well. These problems often reside in basic measurement issues. In fact, most of these problems revolve around various definitions about the measurement of changes, so this is the first thing that requires elaboration.

The following discussion points out several issues related to these concerns, and this chapter eventually leads the reader to the conclusion that it is possible to evaluate changes of either a simple or a complex form using classic structural equation modeling (SEM) techniques (see Horn & McArdle, 1980; Jöreskog & Sörbom, 1979). Most contemporary researchers, however, recognize that they will need to make some key assumptions about the participants, especially those who did not come back a second time (i.e., the incomplete data) and that they will also need to measure multiple indicators of any construct of interest to deal with the important measurement issues. All of this effort must be made just to be sure they can make a reasonably reliable and accurate statement about the changes in the construct. Of course, all of this comes because the trusted models of the analysis of variance (ANOVA) fail in many ways, as is described.

UNIVARIATE MODELS FOR THE MEASUREMENT OF CHANGE

One can formalize change in many ways, but it is relatively simple to look at what are typically known

DOI: 10.1037/14047-012
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

as *gain scores*. For example, one can calculate observed gain scores simply by writing

$$D_n = [Y(2)_n - Y(1)_n], \quad (12.1)$$

where $Y(t)$ is observed at two occasions ($t = 1$ and $t = 2$), and D is used to indicate the changes within a person (where $n = 1 - N$). This approach makes a lot of sense because one can understand this simple calculation of the gain score and the resultant meaning of the gain score: If this score is positive, those interpreting the scores can say the person went up; if this score is negative, they can say the person went down; and if this score is zero, the person did not change at all. The average of these gain scores is interpreted the same way for a group of people.

The variance in the gain should not be ignored, because it is a fundamental clue to the individual differences in the changes. In this simple case of gain scores, the variance of the changes can be written as a simple function of the observed standard deviations ($s[t]$) and the observed correlation over time ($r[1,2]$) as

$$s_D^2 = s(1)^2 + s(2)^2 - 2 * [s(1) * s(2) * r(1,2)]. \quad (12.2)$$

Clearly, the variation in observed change scores can be completely determined from the basic statistics at Time 1 and Time 2. Estimates of the variance of changes become more precise when the two scores are indeed correlated over time. That is, as the correlation is greater than zero (i.e., $r[1,2] > 0$), then the change variability becomes smaller and any difference in means becomes more precise because the second term is subtracted.

One should note that even though the prior calculations (Equations 12.1 and 12.2) are often (and appropriately) termed a *difference score,* largely because of the simple calculation (Equation 12.1), this term is not used in this chapter. Instead, in this discussion the word *difference* is used to refer to *between groups,* or differences between people, and the word *changes* is used to refer to *within group,* or changes within a person (see McArdle, 2009). This approach is typically used when researchers are interested in separating mean differences between some measured groups from the variability of within-group changes in those same groups. Although this wording may seem an unusual restriction of consistent language, it proves helpful as a communication device.

In this nomenclature, if a variable represents the difference between two groups (coded as a dummy variable, $G = 0$ or 1), then a regression model can be written where

$$D_n = \beta_0 + \beta_1 Gn + ed_n, \quad (12.3)$$

so the regression intercept (β_0) indicates the mean of the changes for the group coded 0, the regression slope (β_1) indicates the difference in the changes for the group coded 1, and the regression residual (ed) indicates the residual variance within each group, often conveniently assumed to be the same in both groups. The standard test of whether the regression slope (β_1) is significantly different than zero is identical to the result of the classical within-groups t test. Although the parameters of Equation 12.3 would change if the coding of G (to be, say, $-[1/2]$ and $[1/2]$) is altered, the statistical tests and the essential meaning of the model would not be altered in any way. Indeed, this approach is the formal basis for the standard RANOVA or its two-group analogue, the within-group t test.

A PERSISTENT PROBLEM IN THE MEASUREMENT OF CHANGE

One relatively embarrassing issue is immediately raised in the context of the measurement of change. A classic literature in psychometrics has strongly suggested one not use this simple gain-score calculation because of the notorious unreliability of gain scores (e.g., Cronbach & Furby, 1970). The psychometric reasoning used here is very clear. Assuming a theoretical model for the data at each time for which measurement error is specifically defined can be written as

$$Y(1)_n = y_n + e(1)_n \text{ and } Y(2)_n = y_n + e(2)_n, \quad (12.4)$$

where y is the unobserved true score that is ultimately of interest. In this case, $e(t)$ is an error of measurement at Time t that is not of genuine interest to a researcher. If all this is true, which certainly seems to be the case, then the resulting gain score has the algebraic property

$$\begin{aligned} D_n &= \left[Y(2)_n - Y(1)_n\right] \\ &= \left[y_n + e(2)_n\right] - \left[y_n + e(1)_n\right] \\ &= (y_n - y_n) + [e(2)_n - e(1)_n] \\ &= [e(2)_n - e(1)_n], \end{aligned} \tag{12.5}$$

where the observed gain is simply the ordered difference in the measurement error—something no one is likely to be interested in. It follows that this new gain score simply has no reliable variance, and one writes,

$$\sigma_D^2 = \sigma_e(1)^2 + \sigma_e(2)^2 - 2 * [\sigma_e(1) * \sigma_e(2) * 0]. \tag{12.6}$$

Because this is a model of the population, one writes the deviation as a population parameter (σ_D^2). Cronbach and Furby (1970) conclude that "gain scores are rarely useful, no matter how they may be adjusted or refined. . . . This argument applies not only to changes over time, but also to other differences between two variables" (p. 68). Similarly, Williams and Zimmerman (1996) note that "it might be assumed that the assertion that gain/difference scores are unreliable would be based on empirical studies designed to estimate the reliability of measured gains; however, there is a paucity of data-based investigations" (p. 59). Although not often mentioned, it also seems that any approach that advocates this calculation of gains (e.g., the within-groups *t* test or the RANOVA) must also overcome this unreliability. This statement of the key problem by Cronbach and Furby (1970) was very clear and powerful, so it seemed to have essentially stopped many researchers from doing these kinds of gain-score analyses in developmental psychology. It is interesting, however, that this statement did not seem to keep experimental psychologists from using RANOVA. This kind of statement is still made several decades after it was first advanced.

So with this psychometric result in mind, it may become even more confusing when there exists an equally vocal (but smaller sized) group who are strongly in favor of the use of these observed gain scores (e.g., Nesselroade, 1972; Rogosa & Willett, 1983). Their psychometric reasoning is also clear. The measurement model for this latter approach can be written as

$$Y(1)_n = y(1)_n + e(1)_n \text{ and } Y(2)_n = y(2)_n + e(2)_n, \tag{12.7}$$

so that there is a true score ($y[t]$) and a random noise ($e[t]$) at each time point. If this model were true, then the gain-score calculation is a bit more complex:

$$\begin{aligned} D_n &= [Y(2)_n - Y(1)_n] \\ &= [y(2)_n + e(2)_n] - [y(1)_n + e(1)_n] \\ &= [y(2)_n - y(2)_n] + [e(1)_n - e(1)_n] \\ &= Dy_n + De_n. \end{aligned} \tag{12.8}$$

So, under these assumptions, the calculated gain score is partly signal (Dy) and partly noise (De). This total means that the difference score contains some variance that is signal and some that is noise, and the size of this ratio determines the reliability of the changes. This formulation can also be written as a model of behavior with

$$\begin{aligned} \sigma_D^2 = {} & [\sigma_y(1)^2 + \sigma_e(1)^2] + [\sigma_y(2)^2 + \sigma_e(2)^2] - \\ & 2 * [\sigma_y(1) * \sigma_y(2) * \sigma_y(1,2)]. \end{aligned} \tag{12.9}$$

So now it is clear that the only way changes can be found is if they occur in the true score, which is what is of most interest. Of course, it appears all is not lost in the study of repeated measures because even though there is some noise, the true signal can nevertheless be evaluated, at least in those cases in which the observed gain score is only partly noise.

There is a bit more to this story, of course, including the fact that the advocates of second position are basically graduate students of the advocates of the initial position, but the problematic issue is clear—can the gain in a signal using only two measurements be evaluated? The answer seems to be yes, as long as the true signal actually changes and this change is big enough to outweigh the changes in the random noise.

Dealing With Incomplete Data

A persistent problem with the analysis of change is when some participants have only been measured once. Unfortunately, this situation seems to happen all the time. Obviously, if there is only one measurement (i.e., $Y[1]$), there seems to be no possibility of calculating the observed gains using the simple formula for changes. One easy form of change analysis is to use only those people who came back at both times—this is certainly the logic of complete case analysis, and many researchers seem to believe this

approach is essentially conservative, so it is quite acceptable.

What is not clear is that those who provide a first and a second score may be different in some fundamental way from those who do not return. Of course, it may be as simple as a compliance issue, and given enough time, everyone would participate again. Nevertheless, some form of selection bias can be apparent in the available data. Of course, this is the case in any psychological experiment in which it is clear that not all people asked are likely to participate—even at the first time. In the development of the standard models of analysis (e.g., ANOVA), this factor—a form of external validity—was not a key consideration so now one is forced to use only the data measured at both time points.

In more recent work, however (for review, see Little & Rubin, 1987; McArdle & Woodcock, 1997), it has been very clear that if one is willing to make further assumptions about the selection model, all the available observed data can be used—even data only measured at one point in time. For example, one can combine two groups of participants on the basis of the available data—Group A with both data points and Group B with only the first time point. One can then write a model for the changes for each group as

$$\begin{aligned} \mathrm{A}\left(D_n\right) &= \left\{\mathrm{A}[Y(2)_n] - \mathrm{A}[Y(1)_n]\right\} \text{ and} \\ \mathrm{B}\left(D_n\right) &= \left\{\mathrm{B}[Y^*(2)_n] - \mathrm{B}[Y\left(1\right)_n]\right\}, \text{ so} \\ D^*_n &= [\left(\mathrm{A}D_n\right) : \left(\mathrm{B}D_n\right)], \end{aligned} \tag{12.10}$$

where $Y^*(2)$ indicates that the data are missing and change over time (D^*) is defined to be fully in A and partly in B. The key to the theoretical analysis is the assumption that the statistical information is as if there was only one group, which is equivalent to saying that the change in the second Group B is latent. but its size (mean and covariance) is the same as that in Group A, all of whom are actually measured. The key to the estimation is that the means and covariances at Time 1 are estimated as only one set of numbers.

This newer approach (see McArdle & Nesselroade, 1994) offers a test of invariance over groups—the test here is indexed by a chi-square index that indicates the extent of likely selection bias. That is, at Time 1 it is assumed that the groups may be different, but the results are considered as if all the people came back a second time. Although it is obvious that some of them did not come back, possibly for good reasons (e.g., low scores at Time 1, fatigue), the analysis should describe these people as well. This use of all the available data returns parameter estimates (maximum likelihood estimates) that are neither conservative nor liberal—these estimates are simply considered unbiased and therefore correct. For these reasons, one would use an approach that does deal with incomplete data at every opportunity. That is, one would certainly like to estimate the mean and variance of changes as if no one dropped out and do this in a comparable way across groups, and one would like to provide adequate tests of the assumptions as well.

NOTES ON STRUCTURAL EQUATION MODELING

The use of SEM (see Duncan, 1975; Jöreskog & Sörbom, 1979; Kline, 1998; Loehlin, 2004; McDonald, 1985) offers great flexibility in model creation and model fitting but creates seemingly new problems for both statistical inference and substantive communication. Basically, one can write any model of interest, and then one sees whether this model provides a significantly improved fit over a simpler model, typically using a chi-square index of misfit and the probability of a perfect fit or a close fit to the data.

The basic problem comes with the fitting of multiple alternative models, and many are often listed, although it is very clear that not all of these models are strictly defined in advance of the data collection (see McArdle, 2010). This basically means that the basis of the probability models is not clear. Of course, the knowledgeable reader will note that these seemingly new problems for SEM are actually identical to the standard practices in RANOVA that were never solved. In SEM it is very clear, so as a practical resolution one does not use improper probability values to indicate fit. This chapter gives the reader basic information (i.e., chi-square values and root-mean-square error of approximation misfit indices) and attempts to pull together a compelling and substantively meaningful story about the data.

In turn, this creates many problems for the communication of these basic ideas, and the reader is referred to McArdle and Nesselroade (1994), McDonald and Ho (2002), and Cole and Maxwell (2003) for clearer suggestions on SEM presentations.

ANALYSES FOR MULTIVARIATE DATA OVER TWO OR MORE OCCASIONS

As with dealing with incomplete data, a similar attempt has been made to resolve the persistent problem of random error (see Equation 12.6). McArdle and Nesselroade (1994) first suggested the consideration of a latent change score. In the simplest notation, the basic idea is that if many observed variables are present (say, labeled $W[t]$, $Y[t]$, and $X[t]$) and if one assumes all of these represent a single common factor score (f), then one can consider fitting a model in which multiple variables measured over time are represented as

$$\begin{aligned} W(1)_n &= \lambda w f(1)_n + u w(1)_n \text{ and } W(2)_n \\ &= \lambda w f(2)_n + u w(2)_n \\ X(1)_n &= \lambda x f(1)_n + u x(1)_n \text{ and } X(2)_n \\ &= \lambda x f(2)_n + u x(2)_n \\ Y(1)_n &= \lambda y f(1)_n + u y(1)_n \text{ and } Y(2)_n \\ &= \lambda y f(2)_n + u y(2)_n, \end{aligned} \tag{12.11}$$

under the important assumption that the factor loadings (λm) are identical or invariant at both time points. Of course, given enough data (i.e., three or more variables at each occasion), this rigid assumption of invariant loadings may fail miserably, in which case the factorial description of the tests may need to be altered (see McArdle, 2007).

This general approach basically allows researchers to examine the changes in the latent variable level directly by considering a model where

$$f(2)_n = f(1)_n + \Delta f_n \text{ or } \Delta f_n = [f(2)_n - f(1)_n], \tag{12.12}$$

so the predictor scores in the first part of the equation have an assumed fixed unit regression weight (i.e., =1) so the typical residual is now clearly defined as a latent change score. Most critically, this common latent change score is now theoretically free of errors of measurement.

This means the equation for group differences in changes can be repeated as

$$\Delta f_n = \beta_0 + \beta_1 Gn + ed_n, \tag{12.13}$$

so the regression intercept (β_0) indicates the mean of the latent changes for the group coded 0, the regression slope (β_1) indicates the difference in the latent changes for the group coded 1, and the regression residual (ed) indicates the latent variance within each group (initially assumed to be the same in both groups). The nonstandard test of whether the regression slope (β_1) is significantly different from zero is a latent variable form of the classical within-groups t test.

An alternative possibility that seems simpler to some researchers achieves the same multivariate outcome. First, multiple gain scores (defined by fixed values = 1) are included as

$$\begin{aligned} Dw_n &= \left[W(2)_n - W(1)_n\right] = Dw_n + Dew_n, \\ Dx_n &= \left[X(2)_n - X(1)_n\right] = Dx_n + Dex_n \\ Dy_n &= \left[Y(2)_n - Y(1)_n\right] = Dy_n + Dey_n, \end{aligned} \tag{12.14}$$

where each one is based on some signal and some noise. Second, the random noise can be eliminated by considering the common factor of these observed changes, written as

$$\begin{aligned} Dw_n &= \lambda w \quad f_n + dw_n \\ Dx_n &= \lambda x \quad f_n + dx_n \\ Dy_n &= \lambda y \quad f_n + dy_n. \end{aligned} \tag{12.15}$$

In this approach, the correlation between these multivariate scores can only be due to the latent change score (Δf) multiplied by the respective factor loading (λm). In this case, the invariance of the factor loadings is not evaluated but is simply assumed to be invariant. These are obvious issues that need to be addressed in data analysis.

PLAN OF THE SECOND HALF OF THIS CHAPTER

In what follows, the overall assumption made is that most readers are interested in evaluating change but will actually do so using the classical models of the ANOVA form. This is reasonable, of course, but

these models are limiting in several ways. First, the classical *t* test or ANOVA approach makes it very difficult to go beyond group differences in mean changes. Second, there are very few ways to extend the ANOVA logic to include incomplete data. Third, the multivariate versions of these models (e.g., repeated-measures multivariate analysis of variance; Bock, 1975) allow researchers to form a weighted canonical composite (*c*) to evaluate a multivariate outcome, but this canonical composite is rarely if ever the same as the common factor score (*f*) such researchers may be most interested in. Finally, it is very difficult to extend beyond the simple multivariate analysis of variance models into models based on dynamic concerns (see McArdle & Prindle, 2008).

The data used here come from the well-known Hawaii Family Study of Cognition (HFSC; DeFries et al., 1974; Nagoshi & Johnson, 1994). These analyses begin with an illustration of these methods using classic RANOVA models (O'Brien & Kaiser, 1979) but rapidly move to models based on some relatively recent dynamic structural equation models (from McArdle & Prindle, 2008).

METHOD

Data were chosen from the HFSC to indicate factors of cognitive ability at two points in time, termed Session A and Session C.

Participants

In about 1973, a sampling of more than 6,800 members of more than 1,800 families were measured on at least 15 cognitive tests in Session A. Less well-known is that a sample of 357 of the same people (in more than 115 families) were invited back to take the 15 tests again in 1987 to 1988, and this testing was initially termed Session C. These data were initially collected to estimate the test–rest reliability and changes on each cognitive measure (see Nagoshi & Johnson, 1993, 1994; Nagoshi, Johnson, & Honbo, 1993).

This Session A and Session C sampling was a follow-up of the children from the original sample of the HFSC ($n > 350$), so the reported age range at Session A, 16 to 24 years, was smaller than the overall Session A age range (16–37 years). The follow-up retest was carried out at a 14-year interval from the first testing session. The cognitive tests administered between these two sessions were identical to allow for direct comparison of scores over occasions within people, between people, and over groups. Next, two relatively equal-sized groups were formed on the basis of the reported gender of the participant (male or female).

This analysis focuses on six of the 15 cognitive variables, which were specifically selected to indicate two common factors: (a) a Crystallized Knowledge factor and a Visualization factor of ability (following Horn, 1988). These common factors were indicated in other work by McArdle and Johnson (2004; along with three other common factors) as the best way to organize the outcome variables together in the HFSC. The Crystallized Knowledge factor was indicated by three tests, Vocabulary (VOC), Things (TH), and Word Beginnings and Endings (WBE). In addition, the Visualization factor was indicated by three other tests, Mental Rotation (MR), Card Rotation (CR), and Paper Form Board (PFB).

Obviously, two alternative theories are that (a) all of these six tests are collectively uncorrelated with each other and (b) all of these tests are loaded on the same common factor (see Horn & McArdle, 2007; Spearman, 1904). A more complex hypothesis is that these six tests indeed measure two common factors, but the loadings are not in the correct positions so these two factors should not be labeled Crystallized Knowledge and Visualization. These three alternatives, two more restrictive and one more relaxed, are compared with the specific two-factor models.

Summary statistics for these data are listed in Table 12.1 with means, variance, and correlations split into the two groups. Also highlighted in Table 12.1 are the apparent gender differences in mean levels on some variables.

Basic Models of Mean and Covariance Changes

To study how the participants grew (changed) over time, the results from an RANOVA are listed. This model examines the significance of effects of change over time in the canonical composite using the traditional test level ($p < .05$). By using gender as a

TABLE 12.1

Summary Statistics for Second-Generation Male and Female Participants

Scale	1	2	3	4	5	6	7	8	9	10	11	12
						Male participants (n = 169)						
M	60.3	39.4	24.6	74.7	60.5	44.9	76.5	50.5	47.0	62.7	54.7	44.0
SE	22.4	10.5	8.2	16.4	13.3	13.1	17.5	15.6	15.2	28.1	14.8	13.8
1. VOC1	—											
2. TH1	**.57**	—										
3. WBE1	**.49**	**.41**	—									
4. MR1	.28	.14	.20	—								
5. CR1	.14	.06	.15	.42	—							
6. PFB1	.25	.26	.23	.30	.32	—						
7. VOC2	**.61**	**.32**	**.33**	.13	.08	.15	—					
8. TH2	**.26**	**.38**	**.20**	.00	.07	.25	**.32**	—				
9. WBE2	**.32**	**.17**	**.40**	.14	.10	.14	**.50**	**.33**	—			
10. MR2	.05	−.04	−.03	**.40**	**.25**	**.20**	.26	.20	.17	—		
11. CR2	−.04	−.07	.07	**.30**	**.51**	**.24**	.19	.15	.23	**.62**	—	
12. PFB2	.03	.09	.09	**.26**	**.26**	**.49**	.17	.27	.26	**.42**	**.44**	—
						Female participants (n = 188)						
M	68.8	36.4	28.7	58.2	52.9	44.0	79.6	47.9	50.9	50.9	52.2	42.4
SE	21.9	9.6	8.7	16.7	13.6	15.2	18.5	14.2	14.7	24.6	16.1	14.4
1. VOC1	—											
2. TH1	**.43**	—										
3. WBE1	**.44**	**.36**	—									
4. MR1	.17	.19	.24	—								
5. CR1	.12	.12	.30	**.47**	—							
6. PFB1	.38	.38	.32	**.42**	**.39**	—						
7. VOC2	**.47**	**.30**	**.28**	.13	.12	.13	—					
8. TH2	**.11**	**.51**	**.25**	.08	−.01	.10	**.33**	—				
9. WBE2	**.19**	**.26**	**.35**	.15	.21	.20	**.43**	**.35**	—			
10. MR2	−.04	.16	.01	**.45**	**.40**	**.28**	.19	.14	.28	—		
11. CR2	−.10	.06	.05	**.30**	**.53**	**.17**	.18	.05	.34	**.58**	—	
12. PFB2	.01	.28	.06	**.28**	**.30**	**.39**	.25	.24	.29	**.49**	**.38**	—

Note. Boldface indicates variables factored together between and within time. VOC = Vocabulary, TH = Things, WBE = Word Beginnings and Endings, MR = Mental Rotation, CR = Card Rotation, PFB = Paper Form Board. 1 = Session A; 2 = Session B.

grouping variable, group differences and group differences in canonical changes are also evaluated.

This analysis is intended to be a typical starting analysis. This first analysis focuses on the use of traditional methodological designs to interpret relationships of people as changing, while noting that group differences in how the change occurs may be found.

Models of Factorial Invariance Over Time

One must next focus on building a factor model in which the knowledge outcomes all indicate ability in this one domain as well as for the visual outcomes. With these unobserved factor scores, only the common variation in the outcomes is isolated, which, when one is correct about the model of measurement, provides an improved (i.e., more reliable) score of performance at a given time. The comparison of the two scores over time will then indicate changes. One should test the invariance of the factors to be sure that the structure is unchanged (i.e., invariant) over time and across genders. Without invariance, the direct comparison of latent scores

(both over time and across groups) would not be possible. This two-factor model is compared with a single common factor model of the same data on the six cognitive variables.

Models of Factor Score Changes Over Time

The factor model is then used as the basic component of the cross-lag latent change score regression model. This model uses the factor scores to create a latent change score. The change scores are regressed on the Time 1 factors of the same factor (the lagged effect) and the opposing factor (the crossed effect). This model allows users to make more statements about the effect of how the scores change over time and what influences these changes temporally. Group effects are examined with multiple group analyses.

RESULTS

In this section we present the results of our structural analyses of the HFSC longitudinal data. This begins with an analysis of selection effects and moves to a standard RANOVA-type analysis of mean changes over time. We then consider the assumption of metric factorial invariance over time and use these results in new models of the factor scores over time.

Dealing With Incomplete Data

The data used here illustrate a very large selection effect; as indicated earlier, the full sample size was 6,800 and the retest sample size was about 350. Statistical tests of the differences between the structures of the relationships found in the full data ($N = 6{,}388$) are compared with the retest data ($N = 357$). The results of this analysis are displayed in Table 12.2.

The first model in Table 12.2 started with a fully constrained invariant groups model and constraints were then relaxed with each subsequent model, $\chi^2(39) = 1{,}548$, $\varepsilon_a = .107$. The second model (Model 2) allowed the factor means to vary across groups and indicated modest improvement in fit, $\chi^2(37) = 1{,}408$, $\varepsilon_a = .105$. Next, the factor means were allowed to vary over groups in Model 3 with no gain in model fit, $\chi^2(35) = 1{,}397$, $\varepsilon_a = .107$. Finally, the covariance between the factors was further freed, and there was again no gain in fit in Model 4, $\chi^2(34) = 1{,}388$, $\varepsilon_a = .109$. The last model allowed for unconstrained factor loadings over groups, and this model did not provide evidence of different group structures between the full-data and retest groups, $\chi^2(30) = 1{,}328$, $\varepsilon_a = .113$.

Repeated-Measures Analysis of Variance Models of Mean and Covariance Changes

A doubly multivariate RANOVA (repeated-measures MANOVA) was fit to test the significance of an overall mean difference in multivariate outcomes with Wilks's $\lambda = .025$, with a significance test of $F(6, 347) = 2{,}279$, $p < .001$. This model used the data in which all variables were measured for all people (i.e., complete cases). The overall model fits for Response × Gender effects were indicated by Wilks's $\lambda = .755$, $F(6, 347) = 18.8$, $p < .001$, and for the Response × Time effects, Wilks's $\lambda = .203$, $F(6, 347) = 227$, $p < .001$. The effect of Response × Time × Gender yielded Wilks's $\lambda = .924$, $F(6, 347) = 4.7$, $p < .001$. The between-subjects effect of gender was also significant, $F(1, 352) = 6.1$, $p < .01$.

The next analysis outlined here is a standard RANOVA. The results indicated the relationships of each outcome over time with regard to participant gender. As typically interpreted, over-time effects were found for all variables within subjects except the PFB test: $F_{VOC}(1) = 167$, $p < .001$; $F_{TH}(1) = 237$, $p < .001$; $F_{WBE}(1) = 884$, $p < .001$; $F_{MR}(1) = 55$, $p < .001$; $F_{CR}(1) = 19$, $p < .001$; and $F_{PFB}(1) = 2.6$, $p > .05$. When the Time × Gender interaction was examined, VOC, $F_{VOC}(1) = 6.5$, $p < .05$, and CR, $F_{CR}(1) = 12$, $p < .001$, were considered significant. The effect of gender as a group was again significant for all variables except PFB: $F_{VOC}(1) = 9.7$, $p < .01$; $F_{TH}(1) = 5.6$, $p < .05$; $F_{WBE}(1) = 15$, $p < .001$; $F_{MR}(1) = 53$, $p < .001$; and $F_{CR}(1) = 14$, $p < .001$.

In sum, the standard approach using RANOVA offers many significant results that could be elaborated on. Of course, not all of the results are entirely independent of one another, and the statement of the significance simply in terms of probability is often misleading (see McArdle, 1998, 2010). For these reasons, the probability tests in RANOVA, which are often heralded, may actually be improper estimate of true effects.

TABLE 12.2

Factor Invariance Tests on Visualization and Crystallized Knowledge Comparing Participants Measured Once (*N* = 6,388) With Those Retested (*N* = 357)

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	**FD**	**RT**	**FD**	**RT**	**FD**	**RT**	**FD**	**RT**	**FD**	**RT**
***M* (*SE*)**										
Fgv1	54.3 (0.3)	54.3 (0.3)	**53.8 (0.3)**	**63.3 (0.9)**	53.8 (0.3)	63.3 (0.8)	53.8 (0.3)	63.3 (0.7)	53.6 (0.3)	65.9 (1.1)
Fgc1	68.4 (0.3)	68.4 (0.3)	**68.4 (0.3)**	**68.0 (1.0)**	68.4 (0.3)	68.0 (0.9)	68.4 (0.3)	68.0 (0.9)	68.6 (0.3)	65.2 (1.1)
Variances										
Fgv1	201.2	201.2	197.1	197.1	**199.4**	**150.4**	201.1	126.5	199.7	137.2
Fgc1	228.4	228.4	228.5	228.5	**229.1**	**215.3**	231.2	180.7	232.4	164.7
CovVC	105.6	105.6	105.7	105.7	105.1	105.1	**108.0**	**64.8**	108	63.5
Loadings										
lv1	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00
lv2	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	**0.54**	**0.95**
lv3	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	**0.39**	**0.78**
lc1	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00
lc2	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	**0.88**	**0.95**
lc3	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	**0.75**	**0.78**
Model fit										
χ^2	1,548		1,408		1,397		1,388		1,328	
df	39		37		35		34		30	
RMSEA	.107		.105		.107		.109		.113	

Note. Boldface indicates parameters that have been relaxed across group from the previous model. FD = full data; RT = retest; Fgv1 = visualization factor at Time 1; Fgc = knowledge factor at Time 1; CovVC = covariance of visualization and knowledge factors; lv1 = visual factor loading 1; lv2 = visual factor loading 2; lv3 = visual factor loading 3; lc1 = knowledge loading 1; lc2 = knowledge loading 2; lc3 = knowledge loading 3; RMSEA = root-mean-square error of approximation.

Models of Factorial Invariance Over Time

The next set of models attempt to provide more reliable feedback about the gains within a domain of cognition. Instead of using models within each outcome separately, these analyses try to put the entire set of data together into one model. The invariance tests used here provided evidence for maintaining the same factor structure over time and across groups.

Using the HFSC Session C data from Table 12.1, Table 12.3 displays model tests for invariance for the visual factor that was indicated by mental rotation, card rotation, and paper form board. These three items were fit using a single common factor at each time and across gender groups. A conceptual image of the invariance model is provided in Figure 12.1 for the Visualization factor and Figure 12.2 for the Crystallized Knowledge factor. These represent the parameterization of factor invariance over two time points with a latent change score to model variation in changes over time. Model 1 is based on the complete invariance model, in which all parameters are constant over time and groups, $\chi^2(38) = 231$, $\varepsilon_a = .169$. The next model (Model 2) is one in which the mean constraints over groups were relaxed, and the mean of the factor at Time 1 and the change score were allowed to vary, $\chi^2(36) = 194$, $\varepsilon_a = .157$. In Models 3 and 4, no substantial gains in fit were made by freeing the regression of the change score on the Time 1 factor or the variances of the factor scores progressively, $\chi^2(35) = 194$, $\varepsilon_a = .160$, and $\chi^2(33) = 193$, $\varepsilon_a = .165$, respectively. Model 5 allowed the factor loadings to be free over time, which could provide evidence for a change in outcome weights for the factor over the two groups, $\chi^2(31) = 147$, $\varepsilon_a = .145$.

Testing the invariance of the knowledge factor was done in the same way as the visual factor, and

TABLE 12.3

Visual Factor Invariance Over Two Time Points Within Groups by Male (n = 169) and Female (n = 188) Participants

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
***M* (*SE*)**										
Fgv1	64.49 (1.02)	64.49 (1.02)	**68.9 (1.24)**	**60.3 (1.17)**	68.9 (1.24)	60.3 (1.17)	68.9 (1.21)	60.3 (1.19)	72.9 (1.33)	56.2 (1.26)
lcFgv	5.37 (5.76)	5.37 (5.76)	**1.22 (6.87)**	**4.06 (6.07)**	1.22 (6.87)	4.06 (6.07)	−1.95 (11.1)	6.1 (7.5)	−3.37 (11.8)	6.44 (6.8)
Regression	−0.15 (.09)	−0.15 (.09)	−0.11 (.10)	−0.11 (.10)	**−0.09 (.14)**	**−0.12 (.13)**	−0.06 (.16)	−0.14 (.12)	−0.05 (.16)	−0.15 (.12)
Variances										
Fgv1	127.2	127.2	109.4	109.4	109.4	109.4	**94.7**	**109.4**	106.4	110
lcFgv	124.4	124.4	124.4	124.4	124.3	124.3	**126.3**	**124.3**	147.4	105
Loadings										
l1	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00
l2	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	**0.83**	**0.95**
l3	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	**0.63**	**0.78**
l4	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00
l5	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	**0.83**	**0.95**
l6	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	**0.63**	**0.78**
Residuals										
u1	89	89	90	90	90	90	90	90	92	92
u2	304	304	293	293	292	292	292	292	269	269
u3	142	142	146	146	146	146	146	146	140	140
u4	89	89	90	90	90	90	90	90	92	92
u5	304	304	293	293	292	292	292	292	269	269
u6	142	142	146	146	146	146	146	146	140	140
Model fit										
χ^2	231		194		194		193		147	
df	38		36		35		33		31	
RMSEA	.169		.157		.160		.165		.145	

Note. Boldface indicates parameters that have been relaxed across group from the previous model. Fgv1 = visualization factor at Time 1; lcFgv = latent change score for visualization factor; l1 = loading 1; l2 = loading 2; l3 = loading 3; l4 = loading 4; l5 = loading 5; l6 = loading 6; u1 = uniqueness 1; u2 = uniqueness 2; u3 = uniqueness 3; u4 = uniqueness 4; u5 = uniqueness 5; u6 = uniqueness 6; RMSEA = root-mean-square error of approximation.

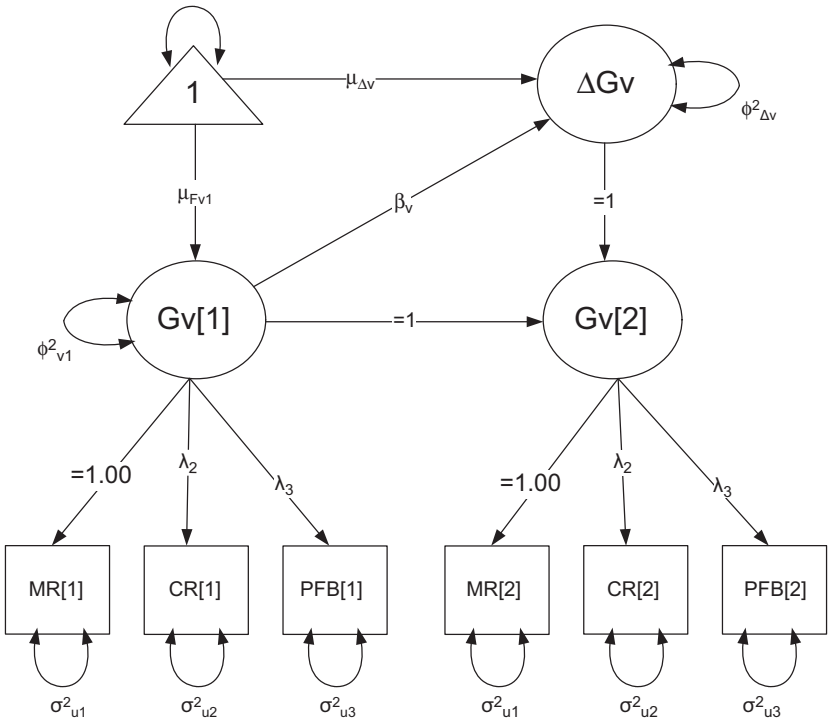

FIGURE 12.1. A latent change score model for Visualization (Gv; with parameters labeled). Gv is indicated by three variables at each time point with equal loadings, and residuals are constrained over time. MR = Mental Rotation; CR = Card Rotation; PFB = Paper Form Board.

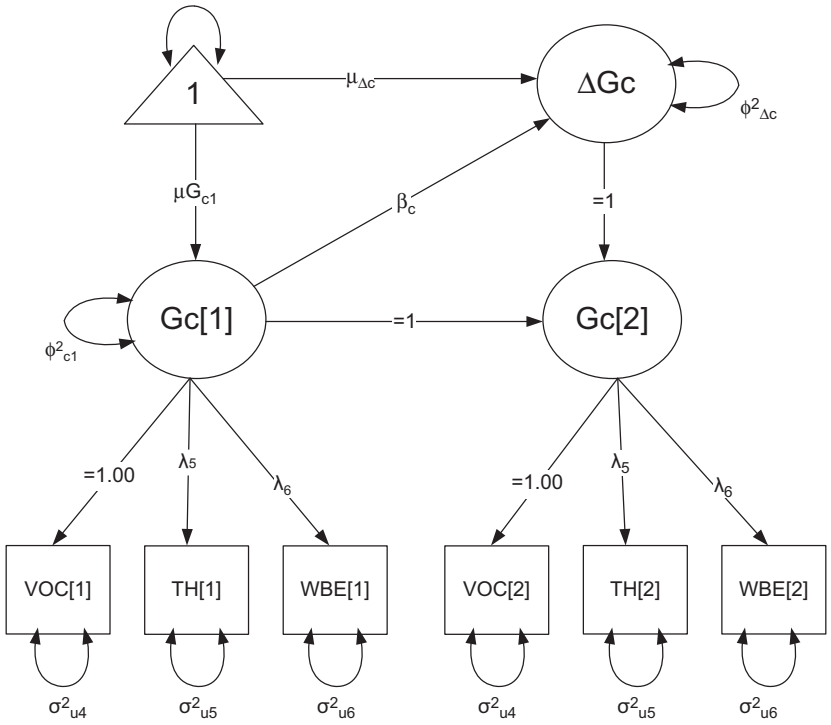

FIGURE 12.2. A latent change score model for Crystallized Intelligence (Gc; with parameters labeled). Gc is indicated by three variables at each time point with equal loadings, and residuals are constrained over time. VOC = Vocabulary; TH = Things; WBE = Word Beginnings and Endings.

the results are outlined in Table 12.4. Progressively relaxed models were fit, with Model 1 being the completely invariant factor model, $\chi^2(38) = 596$, $\varepsilon_a = .287$. Model 2 showed a slight gain in fit by relaxing the means of the factor scores over groups, $\chi^2(36) = 589$, $\varepsilon_a = .294$. Models 3 and 4 then showed no gain in fit by freeing the regression from the first factor score to the latent change score, $\chi^2(35) = 589$, $\varepsilon_a = .298$, and $\chi^2(33) = 588$, $\varepsilon_a = .307$, respectively. Model 5 allowed the factor loadings to be free over groups, $\chi^2(31) = 552$, $\varepsilon_a = .307$.

Models of Factor Score Changes Over Time

The invariant factor models were next coupled together and examined as a cross-lagged model (see Hsiao, 2001) with latent change scores (McArdle, 2009). As indicated earlier, the model that all variables are uncorrelated was tested first and found to have substantial misfit, $\chi^2(132) = 1{,}462$. Then the one-factor model (Figure 12.3) was tested by loading all outcomes onto one factor, $\chi^2(143) = 2{,}298$, $\varepsilon_a = .291$. Next, the two-factor model was tested to increase model fit over the one-factor model, and it appeared to provide poor fit compared with the baseline of no correlation among the outcome variables.

The sequence of model tests for latent changes is displayed in Table 12.5, with model estimates and fit values highlighted. The sequence starts with the most rigid invariance between genders and relaxes the constraints progressively between the gender groups. This model is represented in Figure 12.4 as a structural equation model with latent variables representing the factors and latent change scores. The six measured variables are loaded onto the two factors at both time points with parameters estimated as they are outlined in the model. Model 1 was included as a totally invariant model, with parameters equal over time and between groups. This model did not fit very well, $\chi^2(144) = 967$, $\varepsilon_a = .179$, and was mainly used as a baseline for the subsequent models. Model 2 was fit next, and this model allowed the means of the factor scores to be freed across groups. This model tended to be the best improvement in fit for the previous Crystallized Knowledge and Visualization models, $\chi^2(140) = 909$, $\varepsilon_a = .175$. Model 3 allowed the over-time lagged regressions to be free between groups, but very little gain in fit was observed, $\chi^2(138) = 909$, $\varepsilon_a = .177$.

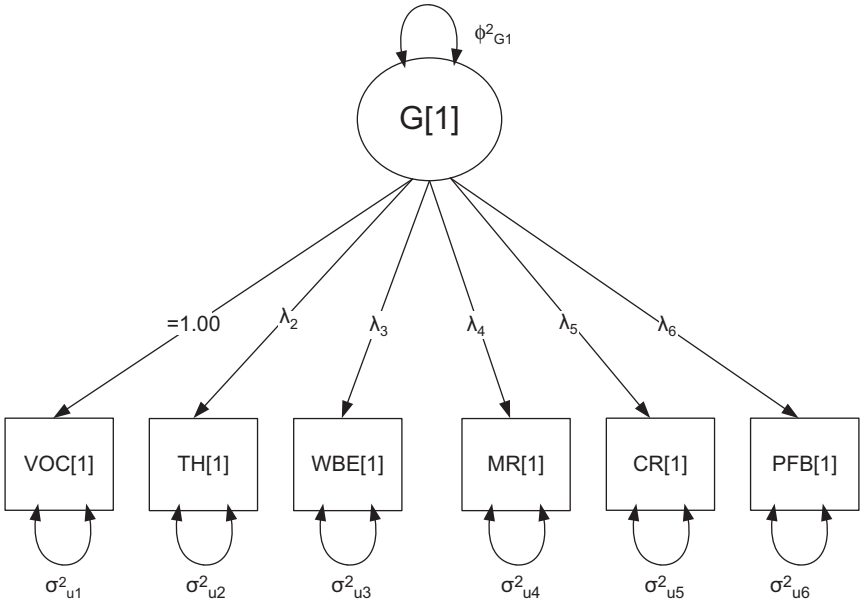

FIGURE 12.3. A general factor model with one factor indicated by six variables per time. The general factor allows each weight to vary, but each item indicates only a general cognitive ability. MR = Mental Rotation; CR = Card Rotation; PFB = Paper Form Board; VOC = Vocabulary; TH = Things; WBE = Word Beginnings and Endings.

TABLE 12.4

Knowledge Invariance Over Two Time Points Within Groups by Male (n = 169) and Female (n = 188) Participants

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
***M* (*SE*)**										
Fgc1	60.9 (1.0)	60.9 (1.0)	**58.7 (1.3)**	**62.8 (1.3)**	58.8 (1.3)	62.8 (1.3)	58.8 (1.4)	62.8 (1.2)	57.7 (1.5)	63.5 (1.3)
lcFgc	31.9 (6.3)	31.9 (6.3)	**31.9 (6.3)**	**30.7 (6.7)**	32.6 (8.0)	30.0 (8.7)	33.6 (7.9)	29.4 (3.0)	32.3 (7.7)	28.9 (10.3)
Regression	−0.20 (.10)	−0.20 (.10)	−0.19 (.11)	−0.19 (.11)	**−0.21 (.14)**	**−0.18 (.14)**	−0.22 (.14)	−0.17 (.16)	−0.22 (.14)	−0.15 (.16)
Variances										
Fgc1	135.1	135.1	131.8	131.8	131.8	131.8	**151.7**	**115.3**	146.3	118.7
lcFgc	101.7	101.7	101.1	101.1	101.0	101.0	**91.7**	**109.9**	91.1	111.4
Loadings										
l1	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00
l2	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	**0.66**	**0.58**
l3	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	**0.54**	**0.55**
l4	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00	= 1.00
l5	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	**0.66**	**0.58**
l6	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	**0.54**	**0.55**
Residuals										
u1	264	264	261	261	261	261	260	260	262	262
u2	108	108	111	111	111	111	110	110	101	101
u3	133	133	132	132	132	132	132	132	130	130
u4	264	264	261	261	261	261	260	260	262	262
u5	108	108	111	111	111	111	110	110	101	101
u6	133	133	132	132	132	132	132	132	130	130
Model fit										
χ^2	596		589		589		588		552	
df	38		36		35		33		31	
RMSEA	.287		.294		.298		.307		.307	

Note. Boldface indicates parameters that have been relaxed across group from the previous model. Fgc1 = knowledge factor at Time 1; lcFgc = latent change score for knowledge factor; l1 = loading 1; l2 = loading 2; l3 = loading 3; l4 = loading 4; l5 = loading 5; l6 = loading 6; u1 = uniqueness 1; u2 = uniqueness 2; u3 = uniqueness 3; u4 = uniqueness 4; u5 = uniqueness 5; u6 = uniqueness 6; RMSEA = root-mean-square error of approximation.

TABLE 12.5

Cross-Lagged Knowledge and Visual Abilities Latent Change Score Model With Groups by Male (n = 169) and Female (n = 188) Participants

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
M **(*SE*)**										
Fgc1	60.9 (1.0)	60.9 (1.0)	**58.6 (1.3)**	**62.9 (1.3)**	58.7 (1.3)	62.9 (1.3)	58.7 (1.3)	62.9 (1.3)	58.6 (1.3)	62.9 (1.3)
lcFgc	36.1 (6.6)	36.1 (6.6)	**38.8 (7.0)**	**35.5 (6.7)**	40.8 (8.5)	33.8 (8.0)	37.1 (9.9)	35.9 (8.6)	35.6 10.2)	36.7 (8.6)
Fgv1	64.5 (1.0)	64.5 (1.0)	**69.0 (1.2)**	**60.3 (1.2)**	69.0 (1.2)	60.3 (1.2)	69.0 (1.2)	60.3 (1.2)	69.0 (1.2)	60.4 (1.2)
lcFgv	16.0 (6.3)	16.0 (6.3)	**7.4 (7.4)**	**15.5 (6.7)**	6.8 (9.8)	15.9 (7.9)	5.1 (10.2)	17.1 (8.3)	6.0 (10.1)	17.2 (8.4)
Regressions										
Fgc → lcFgc	−0.15 (.13)	−0.15 (.13)	−0.09 (.17)	−0.09 (.17)	**−0.12 (.18)**	**−0.06 (.18)**	−0.15 (.19)	−0.03 (.20)	−0.17 (.18)	0.04 (.25)
Fgv → lcFgc	−0.11 (.12)	−0.11 (.12)	−0.19 (.16)	−0.19 (.16)	−0.19 (.16)	−0.19 (.16)	**−0.12 (.20)**	**−0.26 (.20)**	−0.07 (.19)	−0.34 (.25)
Fgc→ lcFgv	0.05 (.12)	0.05 (.12)	0.25 (.18)	0.25 (.18)	**0.26 (.20)**	**0.25 (.19)**	0.29 (.22)	0.24 (.20)	0.22 (.20)	0.31 (.24)
Fgv → lcFgv	−0.39 (.12)	−0.39 (.12)	−0.533 (.16)	−0.533 (.16)	−0.53 (.16)	−0.53 (.16)	**−0.53 (.18)**	**−0.54 (.18)**	−0.46 (.17)	−0.61 (.23)
Covariances										
Δ scores	91.5	91.5	96.0	96.0	96.1	96.1	96.0	96.0	**93.8**	**98.8**
Time 1	60.3	60.3	68.7	68.7	68.7	68.7	68.9	68.9	**58.3**	**78.1**
Variances										
Fc1	136.2	136.2	132.8	132.8	133.0	133.0	133.0	133.0	132.1	132.1
lcFc	95.7	95.7	94.0	94.0	93.6	93.6	93.2	93.2	92.5	92.5
Fv1	123.8	123.8	106.3	106.3	106.3	106.3	106.3	106.3	106.9	106.9
lcFv	104.8	104.8	95.9	95.9	95.8	95.8	95.5	95.5	95.8	95.8
Model fit										
χ^2	967		909		909		908		907	
df	144		140		138		136		134	
RMSEA	.179		.175		.177		.178		.180	

Note. Boldface indicates parameters that have been relaxed across group from the previous model. Fgc1 = knowledge factor at Time 1; lcFgc = latent change knowledge factor; Fgv1 = visual factor at Time 1; lcFgv = latent change visual factor; Fc1 = knowledge factor at Time 1; lcFc = latent change of knowledge factor; Fv1 = visual factor at Time 1; lcFv = latent change of visualization factor; RMSEA = root-mean-square error of approximation.

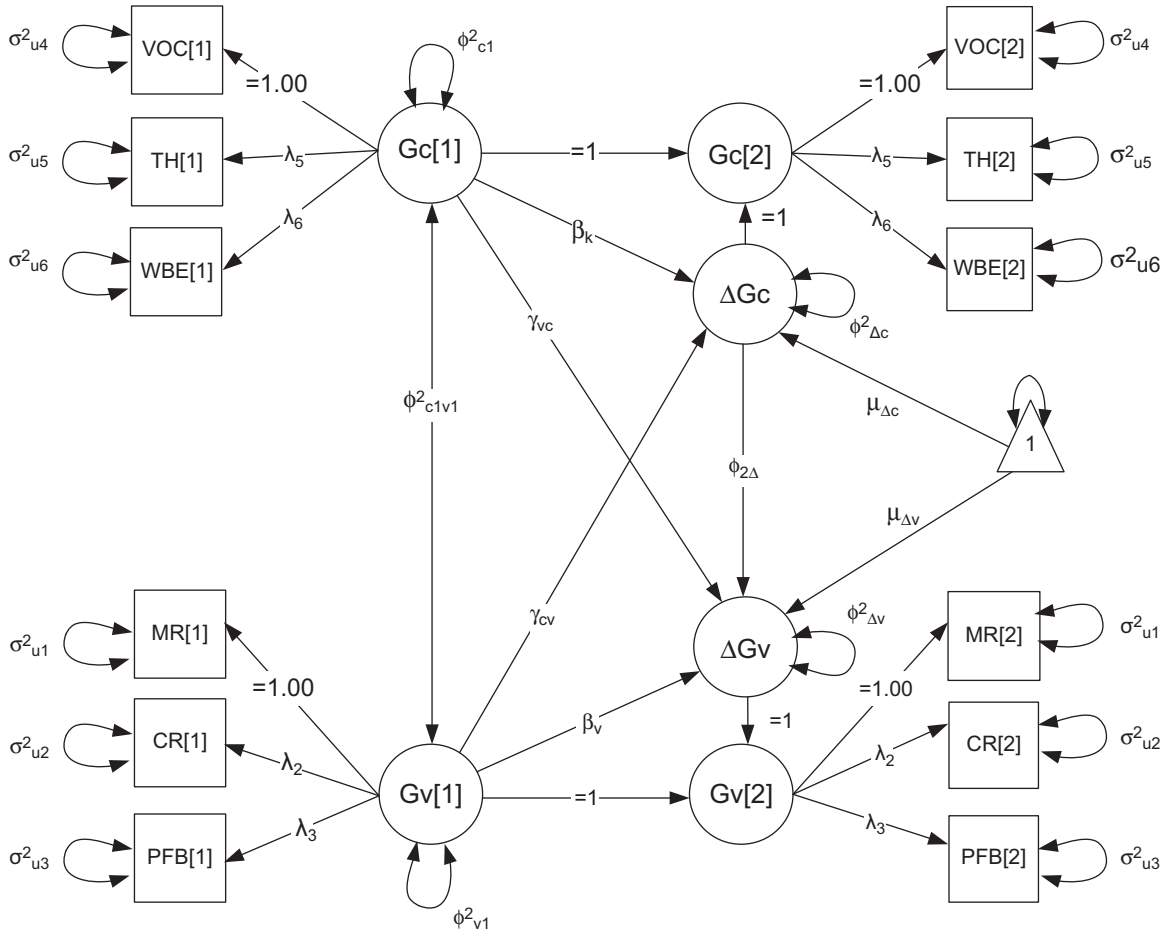

FIGURE 12.4. A cross-lagged latent change score model for Gv and Gc (with parameters labeled). The two latent change score models of Gc and Gv are coupled with regressions from Time 1 to the latent change scores, and the factor scores are allowed to correlate within time. VOC = Vocabulary; TH = Things; WBE = Word Beginnings and Endings; MR = Mental Rotation; CR = Card Rotation; PFB = Paper Form Board.

The next step used here was designed to see whether there was a differential effect of the opposing factors on the latent change scores (Model 4). With little change in model fit, there seemed to be no group differences in the crossed regression weights, $\chi^2(136) = 908$, $\varepsilon_a = .178$. Model 5 was a further attempt to see whether gender differences existed in the relationships between the factor scores. This is a dynamic structural equation model based on the more traditional concept of Group × Time interaction (see McArdle & Prescott, 2010). This approach is performed by allowing the groups to have different covariances between Time 1 latent factors and the latent change scores. In this case, no significant differences in these covariances were found, $\chi^2(134) = 907$, $\varepsilon_a = .180$, so one can conclude that group differences are only in the latent changes in the means over time, and no evidence exists for group interactions.

DISCUSSION

The overall results suggest (a) the two common factors (Crystallized Knowledge and Visualization) represent a good fit to the data, both within people over time and between gender groups; (b) the mean of the latent changes differs over groups, with higher scores at the second time of measurement (at the first time point, there is no latent change score); and (c) the latent changes do not differ between male and female participants.

Although this may seem to be an exhaustive search about different models over time, many other

models of change can be considered. Some of these are simply based on gains changes in proportions (i.e., using logarithms of scores). Others are based on more complex scoring of measurement systems (i.e., Rasch models).

In this chapter, only changes at two specific occasions with specific multiple variables were examined. Several other considerations come into play when one considers models over more time points.

First, there is the possibility that the changes at any occasion are more complex and need to be derived from differences in participants that happened at earlier time points (i.e., two or more lags back). These kinds of changes are not possible to distinguish with only two occasions but can be accomplished with more occasions.

Second, one is likely to consider the wrong interval of the changes. That is, latent changes in these variables could be much different if another range of ages was chosen (young adulthood was used in this chapter). However, if one does not really know the appropriate interval of time, and one hardly ever does (see Gollob & Reichardt, 1987), one needs to experiment with the time lag of this interval (see McArdle & Woodcock, 1997).

Finally, other measurable variables or other common factors may possibly be responsible for the variation in the latent changes (see Shrout, 2010; Sobel, 1995). This problem is considered one of third variables, and this limitation is apparent in any observation study. Of course, if the groups differed on these third variables, one would find some additional group differences of these sizes. However, and in general, variables need to be measured to evaluate their changes.

SUMMARY

The primary purpose of this chapter was to discuss and illustrate models for change using multiple variables and two occasions that extended the options of the principles of the RANOVA. Although the standard RANOVA is still the model of choice of many researchers in psychology, there is obvious room for improvement. Some of these improvements were discussed and illustrated here. In fact, the SEM form of dealing with these kinds of problems shares so many similarities with the prior work that the inclusion of these new measurement models may be well designated as the *new ANOVA*.

Much of what has been presented in this chapter has been anticipated before (e.g., Burr & Nesselroade, 1990; McArdle & Nesselroade, 1994). Although many benefits can accrue from this approach to the measurement of change, it has not yet been put into practice. A first problem is obvious—the computer programs for repeated-measures ANOVA seem to be much easier to use than those for SEM (see Appendix 12.1). However, this is a practical problem that is likely to be reduced in the next generation of computer programs. This implication is that the persistence of teaching the classical methods about group differences seems to be the key problem.

Perhaps teachers have simply not heard about the measurement problems of gain scores or do not realize that the key elements of RANOVA are based on gain scores with little thought given to the measurement requirements. The main reason to carry out all this new SEM work is because it can produce results that are not misleading and are likely to be replicated in the next study. Although the classical RANOVA of group changes can still be considered an important step along the way, now does seem to be the time to embrace the new ANOVA, in which individual differences in changes are also considered. These new approaches are certainly available using current SEM programs (see Appendix 12.1).

APPENDIX 12.1: SELECTED COMPUTER PROGRAM SCRIPTS FOR RUNNING ANALYSES

The procedure for the repeated-measures analysis of variance (RANOVA) is shown first, and the advanced structural equation models are provided with annotations for reference.

SAS RANOVA SCRIPT

```
/* Procedure for RANOVA with six outcomes over
two time points. */
   PROC GLM DATA = datafile;
   CLASS sex; * grouping variable;
```

```
    MODEL p_voc1 p_voc2 p_th1 p_th2 p_wbe1
p_wbe2 p_mr1 p_mr2 p_cr1 p_cr2
    p_pfb1 p_pfb2 = sex / NOUNI;
    REPEATED Response 6 IDENTITY, Time 2;
* identifies number of outcomes and
    how many times repeated;
    LSMEANS sex; * outputs mean scores by sex for
all outcomes at all times;
    RUN; QUIT;
```

MPLUS GV FACTOR SCRIPT

```
    TITLE: HFSC Factor Model - 5
    DATA: FILE = datafile.dat;
    VARIABLE: NAMES = fullid sex
    p_voc1 p_th1 p_wbe1 p_mr1 p_cr1 p_pfb1
    p_voc2 p_th2 p_wbe2 p_mr2 p_cr2 p_pfb2;
    USEV = p_cr1 p_mr1 p_pfb1
    p_cr2 p_mr2 p_pfb2;
    MISSING = .;
    GROUPING = sex (1 = male 2 = female);

    MODEL: ! set up time 1 factor
    Fg1 BY p_mr1@1 p_cr1 p_pfb1;
    Fg1 BY p_cr1 (l1); Fg1 BY p_pfb1 (l3);
    Fg1 (vfg1); [fg1] (mfg1);
    [p_cr1@0 p_mr1@0 p_pfb1@0];

    ! set up time 2 factor
    Fg2 BY p_mr2@1 p_cr2 p_pfb2;
    Fg2 BY p_cr2 (l1); Fg2 BY p_pfb2 (l3);
    Fg2@0; [fg2@0];
    [p_cr2@0 p_mr2@0 p_pfb2@0];

    ! indicator uniquenesses
    p_mr1 (u1); p_cr1 (u2); p_pfb1 (u3);
    p_mr2 (u1); p_cr2 (u2); p_pfb2 (u3);

    ! correlated uniquenesses
    p_mr1 WITH p_mr2; p_cr1 WITH p_cr2;
    p_pfb1 WITH p_pfb2;

    ! latent difference score initialization
    lcfg BY Fg2 @1;
    lcfg ON Fg1 (bfg);
    Fg2 ON Fg1 @1;
    lcfg (vlcfg); [lcfg] (mlcfg);

    MODEL male: ! male model specifications
    Fg1 BY p_mr1@1 p_cr1 p_pfb1;
    Fg1 BY p_cr1 (l1); Fg1 BY p_pfb1 (l3);
    Fg1 (vfg1); [fg1] (mfg1);
    [p_cr1@0 p_mr1@0 p_pfb1@0];

    Fg2 BY p_mr2@1 p_cr2 p_pfb2;
    Fg2 BY p_cr2 (l1); Fg2 BY p_pfb2 (l3);
    Fg2@0; [fg2@0];
    [p_cr2@0 p_mr2@0 p_pfb2@0];

    p_mr1 (u1); p_cr1 (u2); p_pfb1 (u3);
    p_mr2 (u1); p_cr2 (u2); p_pfb2 (u3);

    lcfg BY Fg2 @1;
    lcfg ON Fg1 (bfg);
    Fg2 ON Fg1 @1;
    lcfg (vlcfg); [lcfg] (mlcfg);

    MODEL female: !female model specifications
    Fg1 BY p_mr1@1 p_cr1 p_pfb1;
    Fg1 BY p_cr1 (El1); Fg1 BY p_pfb1 (El3);
    Fg1 (Evfg1); [fg1] (Emfg1);
    [p_cr1@0 p_mr1@0 p_pfb1@0];

    Fg2 BY p_mr2@1 p_cr2 p_pfb2;
    Fg2 BY p_cr2 (El1); Fg2 BY p_pfb2 (El3);
    Fg2@0; [fg2@0];
    [p_cr2@0 p_mr2@0 p_pfb2@0];

    p_mr1 (u1); p_cr1 (u2); p_pfb1 (u3);
    p_mr2 (u1); p_cr2 (u2); p_pfb2 (u3);

    lcfg BY Fg2 @1;
    lcfg ON Fg1 (Ebfg);
    Fg2 ON Fg1 @1;
    lcfg (Evlcfg); [lcfg] (Emlcfg);

OUTPUT: SAMPSTAT STANDARDIZED
TECH4;
```

MPLUS CRYSTALLIZED KNOWLEDE FACTOR SCRIPT

```
    TITLE: HFSC Factor Model - 5
    DATA: FILE = datafile.dat;
    VARIABLE: NAMES = fullid sex
    p_voc1 p_th1 p_wbe1 p_mr1 p_cr1 p_pfb1
    p_voc2 p_th2 p_wbe2 p_mr2 p_cr2 p_pfb2;
    USEV = p_voc1 p_th1 p_wbe1
    p_voc2 p_th2 p_wbe2;
    MISSING = .;
    GROUPING = sex (1 = male 2 = female);
```

```
MODEL: ! set up time 1 factor
Fc1 BY p_voc1@1 p_th1 p_wbe1;
Fc1 BY p_th1 (l3); Fc1 BY p_wbe1 (l4);
Fc1 (vfc1); [fc1] (mfc1);
[p_voc1@0 p_th1@0 p_wbe1@0];

! set up time 2 factor
Fc2 BY p_voc2@1 p_th2 p_wbe2;
Fc2 BY p_th2 (l3); Fc2 BY p_wbe2 (l4);
Fc2 @0; [fc2 @0];
[p_voc2@0 p_th2@0 p_wbe2@0];

! indicator uniquenesses
p_voc1 (u4); p_th1 (u5); p_wbe1 (u6);
p_voc2 (u4); p_th2 (u5); p_wbe2 (u6);
p_voc1 WITH p_voc2; p_th1 WITH p_th2;
p_wbe1 WITH p_wbe2;

! correlated uniquenesses
p_voc1 WITH p_voc2; p_th1 WITH p_th2;
p_wbe1 WITH p_wbe2;

! latent difference score initialization
lcfc BY fc2 @1;
lcfc ON Fc1 (bfc);
Fc2 ON Fc1 @1;
lcfc (vlcfc); [lcfc] (mlcfc);

MODEL male: ! male model specifications
Fc1 BY p_voc1@1 p_th1 p_wbe1;
Fc1 BY p_th1 (l3); Fc1 BY p_wbe1 (l4);
Fc1 (vfc1); [fc1] (mfc1);
[p_voc1@0 p_th1@0 p_wbe1@0];

Fc2 BY p_voc2@1 p_th2 p_wbe2;
Fc2 BY p_th2 (l3); Fc2 BY p_wbe2 (l4);
Fc2 @0; [fc2 @0];
[p_voc2@0 p_th2@0 p_wbe2@0];

p_voc1 (u4); p_th1 (u5); p_wbe1 (u6);
p_voc2 (u4); p_th2 (u5); p_wbe2 (u6);

lcfc BY fc2 @1;
lcfc ON Fc1 (bfc);
Fc2 ON Fc1 @1;
lcfc (vlcfc); [lcfc] (mlcfc);

MODEL female: ! female model specifications
Fc1 BY p_voc1@1 p_th1 p_wbe1;
Fc1 BY p_th1 (El3); Fc1 BY p_wbe1 (El4);
Fc1 (Evfc1); [fc1] (Emfc1);
[p_voc1@0 p_th1@0 p_wbe1@0];

Fc2 BY p_voc2@1 p_th2 p_wbe2;
Fc2 BY p_th2 (El3); Fc2 BY p_wbe2 (El4);
Fc2 @0; [fc2 @0];
[p_voc2@0 p_th2@0 p_wbe2@0];

p_voc1 (u4); p_th1 (u5); p_wbe1 (u6);
p_voc2 (u4); p_th2 (u5); p_wbe2 (u6);

lcfc BY fc2 @1;
lcfc ON Fc1 (Ebfc);
Fc2 ON Fc1 @1;
lcfc (Evlcfc); [lcfc] (Emlcfc);

OUTPUT: SAMPSTAT STANDARDIZED
TECH4;
```

MPLUS BIVARIATE CHANGE SCORE FACTOR MODEL

```
TITLE: HFSC 2 Factor Model - 5
DATA: FILE = hfsc2.dat;
VARIABLE: NAMES = fullid sex
p_voc1 p_th1 p_wbe1 p_mr1 p_cr1 p_pfb1
p_voc2 p_th2 p_wbe2 p_mr2 p_cr2 p_pfb2;
USEV = p_voc1 p_th1 p_wbe1 p_mr1 p_cr1
p_pfb1
p_voc2 p_th2 p_wbe2 p_mr2 p_cr2 p_pfb2;
MISSING = .;
GROUPING = sex (1 = male 2 = female);
ANALYSIS: ITERATIONS = 20000

MODEL: ! Gc model specification
Fc1 BY p_voc1@1 p_th1*.7 p_wbe1*.7;
Fc1 BY p_th1 (l3); Fc1 BY p_wbe1 (l4);
Fc1*15 (vfc1); [fc1] (mfc1);
[p_voc1@0 p_th1@0 p_wbe1@0];

Fc2 BY p_voc2@1 p_th2*.7 p_wbe2*.7;
Fc2 BY p_th2 (l3); Fc2 BY p_wbe2 (l4);
Fc2 @0; [fc2 @0];
[p_voc2@0 p_th2@0 p_wbe2@0];

p_voc1*150 (u4); p_th1*150 (u5);
p_wbe1*150 (u6);
p_voc2*150 (u4); p_th2*150 (u5);
p_wbe2*150 (u6);
p_voc1 WITH p_voc2; p_th1 WITH p_th2;
p_wbe1 WITH p_wbe2;

lcfc BY fc2 @1;
lcfc ON Fc1 (bfc);
```

```
    Fc2 ON Fc1 @1;
    lcfc (vlcfc); [lcfc] (mlcfc);

    ! Gf model specification
    Fg1 BY p_mr1@1 p_cr1*.7 p_pfb1*.7;
    Fg1 BY p_cr1 (l1); Fg1 BY p_pfb1 (l2);
    Fg1*15 (vfg1); [fg1] (mfg1);
    [p_cr1@0 p_mr1@0 p_pfb1@0];

    Fg2 BY p_mr2@1 p_cr2*.7 p_pfb2*.7;
    Fg2 BY p_cr2 (l1); Fg2 BY p_pfb2 (l2);
    Fg2@0; [fg2@0];
    [p_cr2@0 p_mr2@0 p_pfb2@0];

    p_mr1*150 (u1); p_cr1*150 (u2);
p_pfb1*150 (u3);
    p_mr2*150 (u1); p_cr2*150 (u2);
p_pfb2*150 (u3);
    p_mr1 WITH p_mr2; p_cr1 WITH p_cr2; p_pfb1
WITH p_pfb2;

    lcfg BY Fg2 @1;
    lcfg ON Fg1 (bfg);
    Fg2 ON Fg1 @1;
    lcfg (vlcfg); [lcfg] (mlcfg);
    ! combining two LCS models with crosses and
covariances
    Fg1 WITH Fc1 (ct1); lcfg WITH lcfc (cl);
    lcfc on fg1 (glc); lcfg ON fc1 (glg);

    MODEL male: ! male model specifications
    Fc1 BY p_voc1@1 p_th1*.7 p_wbe1*.7;
    Fc1 BY p_th1 (l3); Fc1 BY p_wbe1 (l4);
    Fc1*15 (vfc1); [fc1] (mfc1);
    [p_voc1@0 p_th1@0 p_wbe1@0];

    Fc2 BY p_voc2@1 p_th2*.7 p_wbe2*.7;
    Fc2 BY p_th2 (l3); Fc2 BY p_wbe2 (l4);
    Fc2 @0; [fc2 @0];
    [p_voc2@0 p_th2@0 p_wbe2@0];

    p_voc1*150 (u4); p_th1*150 (u5);
p_wbe1*150 (u6);
    p_voc2*150 (u4); p_th2*150 (u5);
p_wbe2*150 (u6);
    p_voc1 WITH p_voc2; p_th1 WITH p_th2;
p_wbe1 WITH p_wbe2;

    lcfc BY fc2 @1;
    lcfc ON Fc1 (bfc);
    Fc2 ON Fc1 @1;
    lcfc (vlcfc); [lcfc] (mlcfc);

    Fg1 BY p_mr1@1 p_cr1*.7 p_pfb1*.7;
    Fg1 BY p_cr1 (l1); Fg1 BY p_pfb1 (l2);
    Fg1*15 (vfg1); [fg1] (mfg1);
    [p_cr1@0 p_mr1@0 p_pfb1@0];

    Fg2 BY p_mr2@1 p_cr2*.7 p_pfb2*.7;
    Fg2 BY p_cr2 (l1); Fg2 BY p_pfb2 (l2);
    Fg2@0; [fg2@0];
    [p_cr2@0 p_mr2@0 p_pfb2@0];

    p_mr1*150 (u1); p_cr1*150 (u2);
p_pfb1*150 (u3);
    p_mr2*150 (u1); p_cr2*150 (u2);
p_pfb2*150 (u3);
    p_mr1 WITH p_mr2; p_cr1 WITH p_cr2; p_pfb1
WITH p_pfb2;

    lcfg BY Fg2 @1;
    lcfg ON Fg1 (bfg);
    Fg2 ON Fg1 @1;
    lcfg (vlcfg); [lcfg] (mlcfg);

    Fg1 WITH Fc1 (ct1); lcfg WITH lcfc (cl);
    lcfc on fg1 (glc); lcfg ON fc1 (glg);

    MODEL female: ! male model specifications
    Fc1 BY p_voc1@1 p_th1*.7 p_wbe1*.7;
    Fc1 BY p_th1 (l3); Fc1 BY p_wbe1 (l4);
    Fc1*15 (vfc1); [fc1] (Emfc1);
    [p_voc1@0 p_th1@0 p_wbe1@0];

    Fc2 BY p_voc2@1 p_th2*.7 p_wbe2*.7;
    Fc2 BY p_th2 (l3); Fc2 BY p_wbe2 (l4);
    Fc2 @0; [fc2 @0];
    [p_voc2@0 p_th2@0 p_wbe2@0];

    p_voc1*150 (u4); p_th1*150 (u5);
p_wbe1*150 (u6);
    p_voc2*150 (u4); p_th2*150 (u5);
p_wbe2*150 (u6);
    p_voc1 WITH p_voc2; p_th1 WITH p_th2;
p_wbe1 WITH p_wbe2;

    lcfc BY fc2 @1;
    lcfc ON Fc1 (Ebfc);
    Fc2 ON Fc1 @1;
    lcfc (vlcfc); [lcfc] (Emlcfc);

    Fg1 BY p_mr1@1 p_cr1*.7 p_pfb1*.7;
    Fg1 BY p_cr1 (l1); Fg1 BY p_pfb1 (l2);
    Fg1*15 (vfg1); [fg1] (Emfg1);
    [p_cr1@0 p_mr1@0 p_pfb1@0];
```

```
    Fg2 BY p_mr2@1 p_cr2*.7 p_pfb2*.7;
    Fg2 BY p_cr2 (l1); Fg2 BY p_pfb2 (l2);
    Fg2@0; [fg2@0];
    [p_cr2@0 p_mr2@0 p_pfb2@0];

    p_mr1*150 (u1); p_cr1*150 (u2);
p_pfb1*150 (u3);
    p_mr2*150 (u1); p_cr2*150 (u2);
p_pfb2*150 (u3);
    p_mr1 WITH p_mr2; p_cr1 WITH p_cr2;
p_pfb1 WITH p_pfb2;

    lcfg BY Fg2 @1;
    lcfg ON Fg1 (Ebfg);
    Fg2 ON Fg1 @1;
    lcfg (vlcfg); [lcfg] (Emlcfg);

    Fg1 WITH Fc1 (Ect1); lcfg WITH lcfc (Ecl);
    lcfc on fg1 (Eglc); lcfg ON fc1 (Eglg);

OUTPUT: SAMPSTAT STANDARDIZED
TECH4;
```

References

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York, NY: McGraw-Hill.

Burr, J. A., & Nesselroade, J. R. (1990). Change measurement. In A. von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. 1, pp. 3–34). Boston, MA: Academic Press.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558–577. doi:10.1037/0021-843X.112.4.558

Cronbach, L. J., & Furby, L. (1970). How we should measure change—or should we? *Psychological Bulletin, 74*, 68–80. doi:10.1037/h0029382

DeFries, J. C., Vandenberg, S. G., McClearn, G. E., Kuse, A. R., Wilson, J. R., Ashton, G. C., & Johnson, R. C. (1974). Near identity of cognitive structure in two ethnic groups. *Science, 183*, 338–339. doi:10.1126/science.183.4122.338

Duncan, O. D. (1975). *Introduction to structural equation models.* New York, NY: Academic Press.

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development, 58*, 80–92. doi:10.2307/1130293

Horn, J. L. (1988). Thinking about human intelligence. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 645–685). New York, NY: Academic Press. doi:10.1007/978-1-4613-0893-5_19

Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical and statistical model building (MASMOB) in research on aging. In L. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 503–541). Washington, DC: American Psychological Association. doi:10.1037/10050-037

Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100 years* (pp. 205–247). Mahwah, NJ: Erlbaum.

Hsiao, C. (2001). Economic panel data methodology. In N. Snelser & P. Bates (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 4114–4121). Amsterdam, the Netherlands: Elsevier.

Jöreskog, K., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York, NY: Guilford Press.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural analysis* (4th ed.). Mahwah, NJ: Erlbaum.

McArdle, J. J. (1998). Contemporary statistical models for examining test bias. In J. J. McArdle & R. W. Woodcock (Eds.), *Human abilities in theory and practice* (pp. 157–195). Mahwah, NJ: Erlbaum.

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100 years* (pp. 99–130). Mahwah, NJ: Erlbaum.

McArdle, J. J. (2009). Latent variable modeling of longitudinal data. *Annual Review of Psychology, 60*, 577–605. doi:10.1146/annurev.psych.60.110707.163612

McArdle, J. J. (2010). Some ethical issues in factor analysis. In A. Panter & S. Sterber (Eds.), *Quantitative methodology viewed through an ethical lens* (pp. 313–339). Washington, DC: American Psychological Association.

McArdle, J. J., & Johnson, R. C. (2004, October). *Modeling multivariate family data from the Hawaii Family Study of Cognition*. Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Naples, Florida.

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–267). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Prescott, C. A. (2010). Contemporary modeling of gene-by-environment effect in randomized multivariate longitudinal studies. *Perspectives on Psychological Science, 5*, 606–621. doi:10.1177/1745691610383510

McArdle, J. J., & Prindle, J. J. (2008). A latent change score analysis of a randomized clinical trial in reasoning training. *Psychology and Aging, 23,* 702–719. doi:10.1037/a0014349

McArdle, J. J., & Woodcock, J. R. (1997). Expanding test–retest designs to include developmental time-lag components. *Psychological Methods, 2,* 403–435. doi:10.1037/1082-989X.2.4.403

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Erlbaum.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64–82. doi:10.1037/1082-989X.7.1.64

Nagoshi, C. T., & Johnson, R. C. (1993). Familial transmission of cognitive abilities in offspring tested in adolescence and adulthood: A longitudinal study. *Behavior Genetics, 23*, 279–285. doi:10.1007/BF01082467

Nagoshi, C. T., & Johnson, R. C. (1994). Phenotypic assortment versus social homogamy for personality, education, attitudes, and language use. *Personality and Individual Differences, 17*, 755–761. doi:10.1016/0191-8869(94)90044-2

Nagoshi, C. T., Johnson, R. C., & Honbo, K. A. M. (1993). Family background, cognitive abilities, and personality as predictors of education and occupational attainment across two generations. *Journal of Biosocial Science, 25*, 259–276. doi:10.1017/S002193200002054X

Nesselroade, J. R. (1972). Note on the longitudinal factor analysis model. *Psychometrika, 37*, 187–191. doi:10.1007/BF02306776

O'Brien, R. G., & Kaiser, M. K. (1979). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97,* 316–333. doi:10.1037/0033-2909.97.2.316

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*, 335–343. doi:10.1111/j.1745-3984.1983.tb00211.x

Shrout, P. (2010). Integrating causal analysis into psychopathology research (pp. 3–24). In P. E. Shrout, K. M. Keyes, & K. Ornstein (Eds.), *Causality and psychopathology: Finding the determinants of disorders and their cures.* New York, NY: Oxford University Press.

Sobel, M. (1995). Causal inference in the behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York, NY: Plenum Press.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201–293. doi:10.2307/1412107

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*, 59–69. doi:10.1177/014662169602000106

CHAPTER 13

THE STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

Daniel R. Eignor

The *Standards for Educational and Psychological Testing* have been in existence since the mid-1950s. Five versions[1] of the *Standards* have been prepared between then and now, with the most recent, or fifth version, published in 1999. A sixth version of the *Standards* is currently in preparation and should be published in 2012 or 2013. The *Standards* are "joint" in nature in that they were prepared by a joint committee of testing experts representing the three sponsoring organizations (the American Educational Research Association [AERA], the American Psychological Association [APA], and the National Council on Measurement in Education [NCME]).

In this chapter, the purpose of the *Standards* and the instruments for which the *Standards* are expected to apply are discussed, followed by a brief history of the evolution of the versions of the *Standards* over a 50-year period. Finally, a relatively in-depth discussion of the 1999 version of the *Standards,* the version currently in use, is presented. The chapter ends with some concluding comments about the sixth version of the *Standards*, now in preparation.

PURPOSE OF THE *STANDARDS*

Although all versions of the *Standards* have provided a discussion of purpose, albeit in somewhat different ways, providing the purpose statement from the 1999 *Standards,* the version now in use, seems reasonable:

> The purpose of publishing the *Standards* is to provide criteria for the evaluation of tests, testing practices, and the effects of test use. Although the evaluation of a test or testing application should depend heavily on professional judgment, the *Standards* provide a frame of reference to assure that relevant issues are discussed. (AERA, APA, & NCME, p. 2)

The 1999 *Standards* are somewhat less definitive in specifying the measures for which the *Standards* are expected to apply:

> The precise demarcation between those measurement devices used in the fields of educational and psychological testing that do and do not fall within the purview of the *Standards* is difficult to identify. Although the *Standards* applies most directly to standardized measures generally recognized as "tests," such as measures of ability, aptitude, achievement, attitudes, interests, personality, cognitive functioning and mental health, it may also be usefully applied in varying degrees to a broad range of less formal assessment techniques. (AERA et al., 1999, p. 3)

[1]What is being considered as the first version actually consists of two separate documents, one prepared by the American Psychological Association in 1954 and the other prepared by the American Educational Research Association and the National Council on Measurement Used in Education (NCME's original name) in 1955.

DOI: 10.1037/14047-013
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

VERSIONS OF THE *STANDARDS*: SIMILARITIES AND DIFFERENCES

The first set of *Standards,* the *Technical Recommendations for Psychological Tests and Diagnostic Technologies,* was issued by APA in 1954. The next year, AERA and NCMUE (1955) issued a parallel document, *Technical Recommendations for Achievement Tests.* There had been collaboration across organizations in developing these first two documents, and all three organizations agreed that a revision and merging of these documents was needed. In 1963, the first Joint Committee was formed, and this group produced *Standards for Educational and Psychological Tests and Manuals* (APA, AERA, & NCME, 1966), the second version of the *Standards.* All three of these documents essentially focused on the type of information that publishers should provide in test manuals and other publications describing their measures for users and potential users.

Many viewed these initial *Standards* as being too limited in nature, so in 1971 another Joint Committee was formed, and its work produced the 1974 *Standards for Educational and Psychological Tests* (APA, AERA, & NCME, 1974), the third version of the *Standards*. These *Standards* dealt with more than technical documentation of tests and covered the areas of test development, test use, and score reporting.

The first three versions of the *Standards* (APA, 1954; AERA & NCMUE, 1955; APA et al., 1966, 1974) shared a number of features. First, in all these versions, individual standards were labeled as being *essential, very desirable,* or *desirable* in nature. Each standard was, in essence, judged on the basis of its importance and the feasibility of attaining it, and the appropriate label was then attached. Second, all versions listed the standards in a hierarchical fashion. That is, major standards were listed first and then standards that qualified or provided more detail on the major standard were listed. Exhibit 13.1 provides an example of this structure; the standards are taken from the section of the 1974 *Standards* (APA et al., 1974) that dealt with norms and scales. (Note the labeling D5, D5.2, and D5.2.1.) Finally, unique to what is considered the first version of the *Standards* (APA, 1954; AERA & NCMUE, 1955) is that this version contained positive and negative examples of test use and practice. In these examples, the actual names of the tests were listed. This practice was discontinued after the first version was published.

Exhibit 13.1
Example of Hierarchical Structure in 1974 *Standards*

Part D. Norms and Scales

D.5 Derived scales used for reporting scores should be carefully described in the test manual to increase the likelihood of accurate interpretation of scores by both the test interpreter and the examinee. Essential

[Comment: It would be helpful if the number of kinds of derived scales could be reduced to a few with which testers can become familiar. The present variety makes description necessary in each manual. In part the problem is that many different systems are now used that have no logical advantage over others; some may have outlived their usefulness. New scaling methods may be used in attempts to overcome presumed difficulties with older ones. The variety of scales for reporting test scores can create confusion and misinterpretation unless the scale recommended for a given test are clearly and fully explained.]

D5.2 When standard scores are used, the system should be consistent with the purposes for which the test in intended and should be described in detail in the test manual. The reasons for choosing one scale in preference to another should also be made clear in the manual. Very Desirable

D5.2.1 The manual should specify whether standard scores are linear transformations of raw scores or are normalized. Essential

Note. From *Standards for Educational and Psychological Tests,* by American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1974, Washington, DC: American Psychological Association. Copyright 1974 by the American Psychological Association.

The next, or fourth, version of the *Standards* was published by AERA, APA, and NCME in 1985, and the individual standards differed in two ways from those in previous versions. First, instead of labeling individual standards as being *essential, very desirable,* or *desirable,* these descriptors were replaced by *primary, secondary,* or *conditional,*

TABLE 13.1

Names of and Data Contained in the Five Versions of the *Standards*

Version and name	Prepared by	Year	No. of standards	Major standards
1A. *Technical Recommendations for Psychological Tests and Diagnostic Techniques*	APA	1954	163	42
1B. *Technical Recommendations for Achievement Tests*	AERA and NCMUE	1955	111	43
2. *Standards for Educational and Psychological Tests and Manuals*	APA, AERA, and NCME	1966	161	27
3. *Standards for Educational and Psychological Tests*	APA, AERA, and NCME	1974	245	62
4. *Standards for Educational and Psychological Testing*	AERA, APA, and NCME	1985	180	NA
5. *Standards for Educational and Psychological Testing*	AERA, APA, and NCME	1999	264	NA

Note. APA = American Psychological Association; AERA = American Educational Research Association; NCMUE = National Council on Measurements Used in Education; NCME = National Council on Measurement in Education; NA = not applicable.

hence somewhat changing the focus of emphasis for each standard. Second, the hierarchical structure used in the first three versions for labeling standards was replaced by a linear structure, whereby standards within a chapter were simply labeled sequentially (i.e., 2.1, 2.2, 2.3; this structure was also used in the current, fifth version of the *Standards*; AERA et al., 1999).

The fifth version of the *Standards* was published by AERA et al. in 1999, and the one major change in this version was that no categorization or labeling scheme was used to describe the standards. That is, the labels *primary, secondary,* and *conditional* were discontinued. According to the Joint Committee that developed the 1999 *Standards,*

> The present *Standards* continues the tradition of expecting test developers and users to consider all standards before operational use; however, the *Standards* does not continue the practice of designating levels of importance. Instead, the text of each standard, and any accompanying commentary, discusses the conditions under which a standard is relevant. (AERA et al., 1999, p. 2)

Finally, one common feature across all five versions of the *Standards* is the use of comments with individual standards. These comments provide further information to be used in conjunction with the information contained in the standards. Table 13.1 provides a summary of the five versions of the *Standards*, along with some relevant data.

1999 *STANDARDS*

The 1999 *Standards* (AERA et al., 1999) contain 15 chapters separated into three parts. Table 13.2 provides the names of the parts and chapters and the number of standards contained in each chapter. Each chapter begins with an introductory section that does not contain standards, the purpose of which is to provide background material relevant to the chapter. The introductory material in each chapter is followed by the individual standards, and most of the standards also have a comment attached that further explicates what is contained in the standard. Exhibit 13.2 presents two of the standards from the 1999 *Standards* as examples, one from the "Validity" chapter and one from the "Reliability and Errors of Measurement" chapter.

The 1999 *Standards* contain a total of 264 standards (AERA et al., 1999). Table 13.1, previously discussed, contains the number of standards in each published version of the *Standards*. Compared with the number of standards in the 1985 version (AERA et al., 1985), the 1999 *Standards* has 84 more standards. According to Camara and Lane (2006),

TABLE 13.2

Contents of 1999 *Standards*

Part and chapter nos. and titles	No. of standards
I. Test Construction, Evaluation, and Documentation	
1. Validity	24
2. Reliability and Errors of Measurement	20
3. Test Development and Revision	27
4. Scales, Norms, and Score Comparability	21
5. Test Administration, Scoring, and Reporting	16
6. Supporting Documentation for Tests	15
II. Fairness in Testing	
7. Fairness in Testing and Test Use	12
8. The Rights and Responsibilities of Test Takers	13
9. Testing Individuals of Diverse Linguistic Backgrounds	11
10. Testing Individuals with Disabilities	12
III. Testing Applications	
11. The Responsibilities of Test Users	24
12. Psychological Testing and Assessment	20
13. Educational Testing and Assessment	19
14. Testing in Employment and Credentialing	17
15. Testing in Program Evaluation and Public Policy	13

Exhibit 13.2
Examples of Individual Standards in 1999 *Standards*

Chapter 1. Validity

Standard 1.4

If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary.

Comment: Professional judgment is required to evaluate the extent to which existing validity evidence applies in the new situation and to determine what new evidence may be needed. The amount and kinds of new evidence required may be influenced by experience with similar prior test uses or interpretations and by the amount, quality, and relevance of existing data.

Chapter 2. Reliability and Errors of Measurement

Standard 2.15

When a test or combination of measures is used to make categorical decisions, estimates should be provided on the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument.

Comment: When a test or composite is used to make categorical decisions, such as pass/fail, the standard error of measurement at or near the cut score has important implications for the trust- worthiness of these decisions. However, the standard error cannot be translated into the expected percentage of consistent decisions unless assumptions are made about the form of the distributions of measurement errors and true scores. It is preferable that this percentage be estimated directly through the use of a repeated-measurement approach if consistent with the requirements of test security and if adequate samples are available.

Note. From *Standards for Educational and Psychological Testing* (3rd ed.), by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, Washington, DC: American Educational Research Association. Copyright 1999 by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

103 new standards were developed in 1999, and 35 standards from 1985 were eliminated. Other standards were combined. Part of the reason for the increase in the total number of standards in 1999 can be attributed to a decision made by the 1999 Joint Committee to report standards in more than one chapter when possible to do so. The concern was that users of the *Standards* would in many instances refer to only a subset of the total number of chapters and, in the process, miss important standards. Of course, the other reason for an increase in the total number of standards had to do with new developments in the field and the need for these developments to be represented by standards. The advent of computer-administered testing, the existence of test-taker dishonesty enhanced by technology, and the increased importance of educational testing are examples of these new developments.

Note that the configuration of chapters changed from the 1985 *Standards* to the 1999 *Standards*. In two instances, a set of two chapters was merged into a single chapter. Chapters titled "Clinical Testing" and "Test Use in Counseling" in the 1985 *Standards* (AERA et al., 1985) were merged to form part of a single chapter titled "Psychological Testing and Assessment" in the 1999 *Standards* (AERA et al., 1999).

In addition, the 1985 chapters titled "Employment Testing" and "Professional and Occupational Licensure and Certification" were merged to form a single chapter, "Testing in Employment and Credentialing," in the 1999 *Standards*. Finally, a new chapter titled "Fairness in Testing and Test Use" was added to the 1999 *Standards*. A chapter of this sort had not appeared in any of the previous versions of the *Standards*.

The introduction to the 1999 *Standards* (AERA et al., 1999) contains a number of cautions that are important to avoid misinterpretation of the *Standards*. Two of these cautions are particularly important; one has to do with general use of the *Standards* and the other has to do with the specific use of the *Standards* in court litigation, a use that was not initially anticipated:

> Evaluating the acceptability of a test in test applications does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. Specific circumstances affect the importance of individual standards, and individual standards should not be considered in isolation. Therefore, evaluating acceptability involves (a) professional judgment that is based on a knowledge of behavioral science, psychometrics, and the community standards in the professional field to which the tests apply; (b) the degree to which the intent of the standard has been satisfied by the test developer and user; (c) the alternatives that are readily available; and (d) research and experiential evidence regarding feasibility of meeting the standard. (AERA et al., 1999, p. 4)
>
> When tests are at issue in legal proceedings and other venues requiring expert witness testimony it is essential that professional judgment be based on the accepted corpus of knowledge in determining the relevance of particular standards in a given situation. The intent of the *Standards* is to offer guidance for such judgments. (AERA et al., 1999, p. 4)

The reader interested in further discussion of the use of the *Standards* in litigation should consult Sireci and Parker (2006).

2012–2013 *STANDARDS*

Decisions have yet to be made about the final content and chapter structure of the 2012–2013 *Standards* now in preparation. On the basis of a review of a first draft of these *Standards,* it appears that the chapters to be included will be very much like those included in the 1999 *Standards* (AERA et al., 1999). However, if the structure remains the same as it was in the first draft, the three chapters of the 1999 *Standards* that deal with fairness issues—Chapter 7, "Fairness in Testing and Test Use"; Chapter 9, "Testing Individuals of Diverse Linguistic Backgrounds"; and Chapter 10, "Testing Individuals With Disabilities"—would be combined into a single fairness chapter. Finally, in keeping with the 1999 *Standards*, it appears that descriptors such as *primary* will not be associated with individual standards.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing* (Rev. ed.). Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Educational Research Association & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests.* Washington, DC: National Education Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National

Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Camara, W., & Lane, S. (2006). A historical perspective and current views on the *Standards for Educational and Psychological Testing*. *Educational Measurement: Issues and Practice, 25*, 35–41. doi:10.1111/j.1745-3992.2006.00066.x

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice, 25*, 27–34. doi:10.1111/j.1745-3992.2006.00065.x

CHAPTER 14

TECHNICAL REPORTING, DOCUMENTATION, AND THE EVALUATION OF TESTS

Jane Close Conoley, Collie W. Conoley, and Rafael Julio Corvera Hernandez

Previous chapters in Part I of this volume have provided comprehensive coverage of issues critical to test quality, including validity, reliability, sensitivity to change, and the consensus standards associated with psychological and educational testing. In-depth analysis of test quality represents a daunting area of scholarship that many test users may understand only partially. The complexity of modern test construction procedures creates ethical challenges for users who are ultimately responsible for the appropriate use and interpretation of tests they administer. Concerns over the ethical use of tests intensify as the high stakes associated with the results grow.

Test manuals should provide detailed information regarding test score validity and applicability for specific purposes. However, translating the technical information into usable information may be difficult for test developers, who may be more expert at psychometric research than at making their findings accessible to the general reader. Additionally, the information presented in test manuals can be accurate and understandable but insufficient for appropriate evaluation by users.

Fortunately, several sources of review and technical critique are available for most commercially available tests. These critiques are independent of the authors or publishers of the tests. Although test authors and publishers are required by current standards to make vital psychometric information available, experts may be necessary as translators and evaluators of such information. The most well-established source of test reviews is the Buros Institute of Mental Measurements. Established more than 70 years ago by the legendary Oscar Buros, the current Buros Institute continues to publish regular volumes of test reviews and to provide online access to reviews done by experts who often combine psychometric mastery and experience in the particular type of test. Another source of technical information, documentation, and evaluation includes the *Test Critiques* series (Keyser & Sweetland, 1984–2005). Additionally, Psychtests is a Web resource that makes tests, with some description, available to users but lacks extensive evaluation beyond the company's promise that every assessment has been well researched. Details can be found at http://www.psychtests.com/main.

Other resources available to those interested in evaluating the technical information and documentation available on test scores are research studies done to validate the tests or studies that use the particular measures. The *Mental Measurements Yearbooks* provide exhaustive bibliographies of such studies. Users can also investigate textbooks devoted to measurement that provide general information about test standards (e.g., Flanagan, Ortiz, Alfonso, & Mascolo, 2006).

A psychologist or educator in search of an appropriate test product must consider a host of variables before choosing and using a test. This chapter is meant to be a primer that raises the issues a user must consider when choosing a test product and points to useful resources that will inform test-user decision making.

After clarifying the purpose of the assessment process, the initial decision is which type of assessment

DOI: 10.1037/14047-014
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

is required. We begin, therefore, with a brief overview of the types of tests and descriptions of some well-established tests. The descriptions of the tests in this section raise many of the issues that test users must master when choosing testing products. Thus, the subsequent section contains specific questions that must be answered for valid test use.

GENERAL TYPES OF PSYCHOLOGICAL AND EDUCATIONAL TESTS

What follows is not an exhaustive discussion of assessments of each type but rather serves as a general overview of categories in which commonly used psychological and educational assessments can be organized. Examples are offered within each test type. Some tests are listed in more than one category because in many cases the same test can be used for several purposes. All the examples given in this section are commercially available and, thus, represent only a portion of extant assessments. Experimental or basic researchers often use measures of psychological and educational constructs to accomplish their studies. These measures may be copyrighted but could be available in academic journals. They may be individually or group administered and often contain fewer items than commercially published tests. The research measures are typically used for theory building by examining group differences rather than high-stakes decision making about an individual or the future of a program. Even more than when using published tests with accompanying test manuals, the user of noncommercial tests must carefully examine the literature for validity and reliability issues.

Intelligence Tests

Two of the most commonly used intelligence assessments are the Wechsler Intelligence Scales (Wechsler, 1997, 2003) and the Stanford–Binet Intelligence Scales (Roid, 2003). Wechsler's array of tests to measure intelligence were created on the basis of viewing intelligence as a general entity (termed *g*) that characterizes the individual's intellectual capacity as a whole. General intelligence encompasses "an aggregate of specific abilities that are qualitatively different" (Zhu, Weiss, Prifitera, & Coalson, 2004, p. 53). The Wechsler Intelligence Scales, such as the Wechsler Adult Intelligence Scale—Third Edition (Wechsler, 1997) and the Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003), have a strong history of continual development (Groth-Marnat, Gallagher, Hale, & Kaplan, 2000). The Wechsler scales measure abilities such as memory, information-processing speed, abstract reasoning, attention, perceptual organization, verbal comprehension, and quantitative reasoning.

Another set of widely used intelligence tests that measure the two main components of general intelligence (i.e., Spearman's *g*) is Raven's Progressive Matrices and Vocabulary Scales (Raven, Raven, & Court, 2003). Raven's Progressive Matrices measure eductive abilities and the Vocabulary Scales measure reproductive abilities. *Eductive ability* is the ability to think clearly and make sense of complexity; *reproductive ability* refers to the ability to store and recall acquired information (Spearman, 1927). Items on the nonverbal, multiple-choice Progressive Matrices ask the test taker to identify the missing item that completes a pattern.

The Stanford–Binet Intelligence Scales, currently in its fifth edition (Roid, 2003), reflect so-called fluid and crystallized dimensions of intelligence (*Gf* and *Gc,* respectively) on the basis of the work of John Horn and Raymond Cattell (Cattell, 1987; Horn & Cattell, 1966, 1982). Fluid analytical abilities consist of primarily abstract and visual reasoning tasks; crystallized abilities are measured by subtests concerning verbal and quantitative reasoning. The Stanford–Binet Intelligence Scales provide an overall measure of general intelligence constituted with measures of knowledge, fluid reasoning, quantitative reasoning, visuospatial processing, and working memory, nonverbal IQ, and verbal IQ.

The Woodcock–Johnson Psychoeducational Battery—III (Woodcock, McGrew, & Mather, 2001) consists of two conormed batteries, the Test of Achievement and the Test of Cognitive Abilities. The latter is another example of a commonly used instrument to measure intelligence that is based on the *Gf–Gc* theory (Woodcock, 1990).

Another way the Stanford–Binet Intelligence Scales are fundamentally distinct from the Wechsler

Intelligence Scales is that the administration of the Stanford–Binet is designed to be adaptive. On the basis of available information about the examinee, the examiner determines the appropriate starting point and items to administer to limit administration time and maximize the information item responses can provide (Becker, 2003). Because it correlates highly with all other subtests in the fourth edition, the Vocabulary subtest is administered first, enabling the examiner to route to or begin the other subtests at the appropriate difficulty level, based on the examinee's ability. The fifth edition also includes a nonverbal routing test. Other scales, such as the Wechsler Intelligence Scales and the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983), use the examinee's chronological age as a starting point. When a child's chronological age differs from his or her mental age, the Stanford–Binet's adaptive strategy may be particularly advantageous.

Intelligence can also be assessed in a group format. The Armed Services Vocational Aptitude Battery (U.S. Department of Defense, 1984) is considered the "most widespread and important group tests of abilities in existence" (Roberts, Markham, Matthews, & Zeidner, 2005, p. 345). It was initially created as a classification instrument for armed services occupations and thus is not based on a theoretically derived psychological model (Roberts et al., 2005). The Multidimensional Aptitude Battery (Jackson, 1984) is a group-administered paper-and-pencil version of the Wechsler Adult Intelligence Scale—Revised (Wechsler, 1981). The Multidimensional Aptitude Battery measures Performance, Verbal, and Full-Scale IQ. Because it has not routinely been revised as has the Wechsler Adult Intelligence Scale—Revised, however, the Multidimensional Aptitude Battery validity may be suspect (Roberts et al., 2005).

Achievement Tests: Norm and Criterion Referenced

Whereas intelligence tests assess cognitive abilities, achievement tests assess academic accomplishments. Thus, a combination and comparison of such tests are often used to identify learning disabilities, that is, learning difficulties not resulting from low intellectual ability.

Some comprehensive achievement assessments assess a wide range of academic abilities, and others are used as brief or screening instruments (Flanagan et al., 2006). Examples of comprehensive achievement batteries are the Wechsler Individual Achievement Test (2nd ed.; Wechsler, 2002), and the Woodcock–Johnson—III Tests of Achievement (Woodcock et al., 2001). Brief or screening instruments do not provide the depth of the longer comprehensive tests (Flanagan et al., 2006). Examples of brief assessments are the Hammill Multiability Achievement Test (Hammill, Hresko, Ammer, Cronin, & Quinby, 1998), the Wide Range Achievement Test—3, and the Young Children's Achievement Test (Hresko, Peak, Herron, & Bridges, 2000).

A third category of academic achievement instruments includes tests designed to measure a particular academic skill (e.g., mathematics) or specific processing ability underlying the development of a specific academic skill (Flanagan et al., 2006). The Woodcock–Johnson—III Diagnostic Reading Battery and the Comprehensive Mathematical Abilities Test are examples of instruments designed to measure specific academic skills.

The Woodcock–Johnson Psycho-Educational Battery—Revised (Woodcock & Johnson, 1989) is a popular academic achievement assessment and arguably the most comprehensive academic battery currently in use (Johnstone, Holland, & Larimore, 2000). The nine individual subtests that make up the Standard Battery require 50 to 60 minutes to administer. The subtests evaluate the following abilities: recognition of letters and words; reading comprehension; performance of various mathematical calculations; solving practical math problems; diction; basic quality of written expression; and the general knowledge of science (biological and physical), geography, government, economics, arts, music, and literature.

An important distinction between the uses of achievement tests lies in the purpose of the results. Achievement tests can use two different types of interpretation, norm-referenced interpretation and criterion-referenced interpretation. Tests used to rank students in relation to others use a method of norm-referenced interpretation. *Norm-referenced interpretations* are statements such as "She performed

second highest in her class of 28 students" or "She performed in the top 10% of the students who have taken this test." In contrast, criterion-referenced interpretation tests identify the specific learning tasks that a student can or cannot perform. An example of a criterion-referenced interpretation is "She understood mathematics up to algebra." Standardized achievement tests are widely used to assess the performance of schools via the norm-referenced interpretation. However, both types of interpretation could be obtained from the same assessment if the tester has a deep understanding of the test content and the characteristics of the norm group (Gronlund, 2003).

Personality Assessments: Structured and Projective

Personality assessments in psychology serve five purposes: identifying psychopathology and diagnoses, describing and predicting everyday behavior, informing psychological treatment, monitoring treatment changes, and using psychological assessment as a treatment (Smith & Archer, 2008). Personality tests can be categorized into three types: *performance based* (also known as *projective*), *self-report* (often referred to as *objective*), or *behavioral*. Performance-based personality assessments require the test taker to respond to a stimulus in an unstructured format so that important individual characteristics may emerge (Smith & Archer, 2008). The Rorschach Inkblot Test (Exner, 2003) and Thematic Apperception Test (Cramer, 1996) are examples of performance-based or projective personality assessments.

The self-report or objective personality assessments provide respondents with a structured response format that provides several response choices to a number of questions about themselves. These types of assessments can be further divided into omnibus (also known as broad-band) or narrow-band assessments. The narrow-band assessments focus on a few personality characteristics in depth, whereas the omnibus measures assess multiple domains (Smith & Archer, 2008). The Personality Assessment Inventory (Morey, 1991), an omnibus self-report personality measure, assesses for a variety of constructs, including depression, anxiety, thought disorder, and drug abuse. An example of a narrow-band self-report personality measure is the Rosenberg Self-Esteem Scale (Rosenberg, 1965) because it measures only self-esteem.

Vocational Assessments

The three assessment areas of vocational interests, needs and values, and abilities make up vocational assessments that inform career counseling (Watkins, Campbell, & Nieberding, 1994). Of these constructs, vocational interest is the most frequently assessed (Hansen, 2005). Interest inventory profiles help the career counselor develop hypotheses about clients that guide career exploration and inform a person's occupational and educational decisions through a better understanding of personal interests and career possibilities.

The three most frequently used interest inventories incorporate John Holland's (1997) vocational choice theory, which specifies six vocational interest personality types: realistic, investigative, artistic, social, enterprising, and conventional (Holland, 1997). The most popular vocational inventories are the Self-Directed Search (Holland, 1985), Strong Interest Inventory (Harmon, Hansen, Borgen, & Hammer, 1994), and the Campbell Interest and Skill Survey (Campbell, Hyne, & Nilsen, 1992). The estimation of a person's ability to perform tasks related to his or her interests is measured in different ways among these inventories. Although the Campbell Interest and Skill Survey includes Skill scales that parallel all 98 interests on its Interest scale, the Strong Interest Inventory has a companion instrument called the Skills Confidence Inventory (Betz, Borgen, & Harmon, 1996) that provides ability estimates for Holland's six types.

Popular interest inventories often used with high school students are the Kuder Occupational Interest Survey (Kuder & Zytowski, 1991) and the American College Testing Interest Inventory (American College Testing Program, 1995). A common inventory for people considering nonprofessional career choices is the Career Assessment Inventory (Johansson, 1986).

Measures of work values, that is, what people want and expect from work (Nord, Brief, Atieh, & Doherty, 1990), provide information related to the

vocational issues of career choice, job satisfaction, and motivation. The Minnesota Importance Questionnaire (Gay, Weiss, Hendel, Dawis, & Lofquist, 1971) measures vocational needs and work-related values to help identify preferences that are important for making career choices. The O*NET is a database designed to update and expand the Minnesota Importance Questionnaire's occupational information (Rounds & Armstrong, 2005). The Work Importance Profiler (U.S. Department of Labor, 2002) is a computerized assessment of values based on the O*NET. On completion of the Work Importance Profiler, the computerized assessment provides a list of occupations that match the individual's values profile.

A shorter pencil-and-paper measure of work values is the Work Importance Locator (U.S. Department of Labor, 2000). The Work Importance Locator uses a card-sorting task of work needs based on the Minnesota Importance Questionnaire. Each card has a work need printed on it that a person places in order of importance. This card-sorting procedure has lower reliability than the computerized Work Importance Profiler and Minnesota Importance Questionnaire.

The Values Scale and the Salience Inventory (Nevill & Super, 1986b, 1986a) are two values assessments based on Super's (1980) vocational theory and developed through a multinational effort (Ferreira Marques & Miranda, 1995). The Values Scale (Nevill & Super, 1986b) creates a hierarchy of 21 values based on the relative scores obtained on each of the 21 five-item scales. The Salience Inventory (Nevill & Super, 1986a) assesses the importance of life and career goals (i.e., home and family, working, studying, leisure activities, community service). Items probe participation, commitment, and value expectation.

Neuropsychological Assessments

Neuropsychology explains human functioning via the relationship between the brain and human behavior. Neuropsychological tests strive to provide "a clear and coherent description of the impact that brain dysfunction has had on a person's cognitions, personality, emotions, interpersonal relationships, vocational functioning, educational potential, and ability to enjoy life" (Groth-Marnat, 2000, p. 3). Neuropsychological assessment can be used for many purposes in medical, law, education, and research settings (Hebben & Milberg, 2009). Interviews, standardized scale tests, and questionnaires allow the intensive study of the brain–behavior relationship by neuropsychologists (Lezak, Loring, & Howieson, 2004). Neuropsychological tests frequently assess the domains of learning and memory, language functions and academic skills, attention, mental activities, visuoconstructive abilities, and emotional functioning.

Learning and memory. Word list learning tasks, such as the Rey Auditory Verbal Learning Test (Rey, 1964) and the first and second editions of the California Verbal Learning Test (Delis, Kramer, Kaplan, & Ober, 1987, 2000), assess learning and verbal memory by asking the test taker to recall words from a list after hearing or reading them over several trials. The Rey–Osterrieth Complex Figure Test also assesses visual memory, visual perception, drawing, and constructional praxis (Moye, 1997). Lower scores do not simply reflect deficits in visual memory because visuoperceptual and visuoconstructive ability impairments could also affect performance (Janowsky & Thomas-Thrapp, 1993). The Taylor Complex Figure is an alternative form of the Rey–Osterrieth Complex Figure Test that can be used when an individual needs to be retested. Other assessments of visuospatial learning and memory include the Continuous Visual Memory Test (Trahan & Larrabee, 1988) and the Visual Object Learning Test (Glahn, Gur, Ragland, Censits, & Gur, 1997). The Visual Object Learning Test is a computer-administered, object-list learning task including learning trials, a distraction trial, and immediate and delayed recall trials. The Continuous Visual Memory Test assesses recognition memory and distinguishes between impairments in visual memory and visual perception. It involves a visual discrimination task, an acquisition task, and a delayed recognition task.

The original and revised Wechsler Memory Scales are the most commonly used assessments of memory in clinical and neuropsychological practice (Franzen & Iverson, 2000). The original Wechsler Memory Scale (Wechsler, 1945) evaluated memory

as a unitary construct. Modern conceptualizations of memory include distinctions between short-term (primary memory store) and long-term (secondary memory store) memory, as articulated in the Atkinson–Schiffrin (1968) model. Another distinction made is that between procedural (or implicit) and declarative, episodic, or explicit memory (Tulving, 1985). Procedural memory is typically measured with speed of learning or shifts in choice bias (Helmes, 2000). Thus, current practice has suggested the use of a battery of tests because memory is conceptualized as "differentiable into stages and modalities, and recall versus recognition" (Franzen & Iverson, 2000, p. 195).

Language. The Wechsler Adult Intelligence Scale—Third Edition (Wechsler, 1997) Verbal Intelligence subtests are commonly used to evaluate general verbal skills. The scale's subtests relevant to language function include Information (range of knowledge), Comprehension (judgment), and Vocabulary (vocabulary level). The Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1976) is commonly used among psychologists as a brief screening for expressive language abilities by having individuals spontaneously identify 60 pictures of objects ranging from easily identifiable to relatively unfamiliar. Individuals are given a 20-second time frame to name each picture. The Controlled Oral Word Association Test (Benton, Hamsher, & Sivan, 1994), commonly referred to as the Word Fluency Test or the FAS Test, measures the ability to spontaneously state words beginning with a certain letter (usually *F, A,* or *S,* hence the nickname) or belonging to a specific category (e.g., animals or foods). This test is also structured around a limited time frame in which individuals produce as many words as possible that fit the given criteria.

Neuropsychological batteries. A core battery of tests typically constitutes a neuropsychological assessment. The batteries can be a formally developed set of subtests within a larger test or consist of the neuropsychologist's preferred tests (Groth-Marnat, 2000). The major formally developed test batteries are Wechsler Intelligence Scales, Wechsler Memory Scales, the Halstead–Reitan Neuropsychological Test Battery (Reitan & Wolfson, 1993), the Luria-Nebraska Neuropsychological Battery (Golden, Purisch, & Hammeke, 1985), and the Boston Process Approach (Kaplan, 1988). The utility of the Wechsler Intelligence and Memory Scales in neuropsychological assessment extends beyond the evaluation of intelligence and memory into the functional consequences of cognitive impairment and brain damage (Groth-Marnat et al., 2000). Each approach has various advantages and disadvantages that assist in guiding the selection of the batteries for different situations.

Hebben and Milberg (2009) suggested the following distinguishing qualities differentiating the latter three batteries. The Halstead–Reitan Neuropsychological Test Battery has a wealth of validating data, demonstrated good reliability across different patient groups, and can be administered by a technician. However, it is lengthy and inefficient, made up of complex measures and, probably for these reasons, declining in popularity. The battery requires extensive training for proper administration and interpretation. In contrast, the Luria-Nebraska Neuropsychological Battery takes less time to administer and has single scales for various functional and cognitive domains. The Luria-Nebraska Neuropsychological Battery is also declining in popularity, however, and the Boston Process Approach remains the most popular. The flexibility of the Boston Process Approach allows for matching of the tests administered to the specific referral question. The drawbacks include a dependence on observational skills for its use, necessity of specific training, and a limited set of normative data for qualitative results.

Behavior Assessments

Many forms of psychological assessment focus on what the examinee has (i.e., character traits, attributes), but behavioral assessment emphasizes what the examinee does (Ramsay, Reynolds, & Kamphaus, 2002). In other words, behavioral assessments emphasize situational determinants of behavior and focus on antecedents and consequences rather than underlying traits (Groth-Marnat, 2009). Fortunately, behaviors can be measured in analogue or naturalistic settings, and assessment is concerned with both overt and covert behaviors. Clearly observable overt behaviors have

been the historical focus of behavioral assessment; however, the rising influence of the cognitive–behavioral orientation ushered in the inclusion of covert behaviors, such as thoughts, desires, and feelings (Ramsay et al., 2002).

Behavioral assessment includes a wide variety of strategies that usually measure the existence and frequency of behaviors. Strategies are generally categorized as behavioral interviewing, behavioral observation, behavioral rating scales, cognitive–behavioral assessment, recording cognitions, psychophysiological assessment, and self-report inventories (Cone, 1978; Groth-Marnat, 2009). Methods for the assessment of behavior are also categorized on a continuum ranging from direct to indirect (Kratochwill, Sheridan, Carlson, & Lasecki, 1999). An assessment's place on the continuum is determined by the extent to which the assessment measures the clinically relevant behavior and the extent to which the assessment measures the behavior in its naturally occurring time and place (Cone, 1978). Direct assessment methods measure clinically relevant behavior in its natural context (time and place; Shapiro & Kratochwill, 2000). Self-monitoring, analogue assessment, naturalistic observation, and counting discrete behavioral events are examples of direct procedures. Interviews, self-reports, and ratings by others are indirect methods of behavioral assessment.

Adaptive behavior is "the performance of the daily activities that are required for social and personal sufficiency" (Sparrow, Cicchetti, & Balla, 2005, p. 6). Scales for the purpose of measuring adaptive behavior have played a central role in the diagnosis of developmental disability (DeStefano & Thompson, 1990). The criteria used by the American Association on Intellectual and Developmental Disabilities to determine the presence of intellectual disability are deficits in intellectual functioning and adaptive behavior (Luckasson et al., 2002). Adaptive behavior scales measure skills that are essential to a child's ability to function successfully in a variety of environments. Because the results identify specific skills that a child has not acquired, the assessments provide valuable clinical information for designing interventions. The Vineland Adaptive Behavior Scales (2nd ed.; Sparrow et al., 2006), is a popular tool for the assessment and diagnosis of developmental disability. The Adaptive Behavior Assessment System (2nd ed.; Harrison & Oakland, 2003), is an assessment closely based on the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1994) and American Association on Intellectual and Developmental Disabilities (Luckasson et al., 2002) criteria for developmental disability.

INVESTIGATING TESTS' TECHNICAL QUALITIES AND DOCUMENTATION

The previous section provided a dense review of the major types of tests and examples of the most commonly used tests within each type. The test descriptions also included features of the tests that require special inquiry from potential users to ensure valid use. Note that the tests were described with full names, authors, related research, norm groups, test administration strategies, theoretical bases, and so on. These finer points of test qualities are expanded on in the following section. Users must seek out documentation of these technical features if they are to use the test with confidence.

What Score, What Strategy, Which Audience, What Use

The sections that follow provide a list of questions that test users should consider as they choose tests. Careful analyses of each of these variables—scores, testing strategies, audience, and ultimate use—are critical for valid test use.

What score? The search for a test usually begins with the user identifying the variables of interest. These variables are best represented by the scores that are derived from various tests. For example, a researcher might want to know about locus of control or extraversion. An educator may want to measure children's reading fluency and comprehension. Although test titles may contain the words *personality* or *reading,* these are gross descriptions that must be examined closely to be sure that the chosen test actually delivers the score of interest. For example, many personality tests offer extraversion scores but

not locus of control estimates. Similarly, many reading tests have a comprehension scale but not a fluency measure.

Researching the exact meaning of a construct measured by a test is too often overlooked. Marsh (1994) coined the term *jingle-jangle fallacy*, which cautioned that the jingle assumes that two scales with the same name measure the same construct and that the jangle assumes that two scales with different names measure different constructs. The confusion of constructs with similar names occurs, for example, with the measurement of achievement goals. Hulleman, Schrager, Bodman, and Harackiewicz (2010) found an assortment of different meanings as they performed a meta-analysis to understand the meaning of achievement goals—for example, performance versus mastery—in the literature. The importance of reading the test items and the literature associated with a measure cannot be overemphasized.

The *Mental Measurement Yearbooks* (e.g., Spies, Carlson, & Geisinger, 2010) facilitate this search via their score index. Test users can begin their search for appropriate assessments by finding tests associated with the scores that interest them. Of course, the name of a score, although important, is just a name until the score is related to the construct of interest. The test user must find information that gives the score credibility as a valid measurement of a particular construct. The key issue is, of course, validity. Does the score really provide information about what it purports to measure?

Determining the validity of the score requires expert assessment of the research surrounding the test. Have predictive validity analyses or multitrait, multimethod analyses been performed? Have the test authors attended to reliability issues associated with repeated testing or tests given over a particular time period? If the scores are based on responses to a variety of items (scales), do those scales show strong within-scale correlations of items while showing adequate between-scale differences? That is, does the author report a robust factor structure for the measure? If not, differential score reporting may be compromised because the test may, for example, measure some general aspect of intelligence but give little or no reliable information about constituent components of intelligence (e.g., memory, attention, or information-processing speed).

What strategies? A variety of tests that derive identical (or highly correlated) scores may differ dramatically in how they are administered. Some are individually administered; some are timed; some require significant equipment and training to implement. Others can be taken at computer terminals and scored electronically or given to large groups. A test user may have time, economic, and implementation constraints that favor one test over another. A common feature to be considered is whether the test can be administered in a group or must be done individually. If a group test provides a measure of intelligence that meets the user's needs, it may be the economical choice in terms of time, money, and convenience. In contrast, if another user must be certain that the people tested performed at their peak effort, individual administration may be necessary. Some tests take hours to complete, whereas others can be finished in minutes—preferable, perhaps, if initial screening is the user's goal.

The administration requirements among tests may require some trade-offs. For example, most researchers have more confidence in individually administered intelligence tests, but if a test user wants a general sense of cognitive functioning, a group-administered test may suffice. This choice could preclude using the score as evidence of the need for special services in schools, but the convenience of the group measure may allow for general screening followed by more targeted assessment for smaller groups whose scores indicate exceptionally low or high functioning.

What audience? Tests are developed for certain audiences both in their administration strategies and in their interpretive strengths. A common discriminator among tests is recommended age range. If the test user wants to assess 3-year-olds, it is vital that the test be designed to be compelling to preschool-aged children and contain norms that include 3-year-olds. Another aspect of a test audience is its general intellectual capacity. That is, if a test user plans to assess a population of individuals with very serious cognitive delays, the test must be designed with enough lower level items to allow for

discriminations to be made within that population. This quality is often referred to as the *low floor* of the test. Conversely, if the testing is aimed at highly gifted individuals, the test must contain multiple items with a high degree of difficulty—the so-called *high ceiling* of the test. Documentation should be available that allows for the test user to determine whether a particular test is appropriate for general populations, exceptional populations, or both. These data would appear in norm group descriptions and analyses.

Gender, social class, geographic regions, and ethnicity constitute other common audience characteristics. Test users need evidence that the groups used for the test norms included individuals such as the intended test takers, or they should be ready to interpret the scores with special care. For example, if the test's interpretive norms were based on samples of individuals who were all male or lived in only one region of the United States or were not members of various ethnic groups, test users should view the published norms with skepticism as a basis for score interpretation unless the user's population of test takers is similar. If the test user's population is homogeneous and the test's norm group contains the user's population, then the user can be reassured if the test manual provides information that the user's population does not differ from the general population's norm information. Some tests provide separate norms for each subgroup if the norms vary between groups. If the group a test user wants to assess is not included in the test's norm information, further research would be necessary to confirm that scores did not show a systematic difference on the basis of ethnicity, region, or social class in ways that added error to measuring the variable of interest.

What use? The scores derived from tests may be used for a wide variety of purposes. As described in the first section of this chapter, test scores may enable services for individuals (e.g., access to gifted and talented education), determine drug regimens (e.g., to reduce thought disorder), or determine whether a person achieves a professional position (e.g., job-related math achievement). Some tests are considered to be high-stakes assessments because results have significant, material effects on individuals or organizations. Current state kindergarten-through-12th-grade testing systems are high stakes because they determine the ranking of schools within a school district and can be used as the basis for closing or restructuring schools that fail to meet established cut scores. Obviously, high-stakes tests require substantial validity and reliability evidence before use. Test users should seek evidence that scores are valid and reliable markers of the test constructs. School districts, for example, should require evidence that achievement tests used to measure yearly academic progress are, in fact, sensitive to change for an individual across time to allow cognitive growth interpretations and that they are excellent representations of the content standards associated with each grade.

Developing such evidence is not a trivial undertaking because it requires giving forms of the tests to representative samples of children, relating the scores achieved to other measures known to represent achievement, checking for irrelevant sources of error resulting from item type or time restrictions, and so on. Evaluating and interpreting reports of these efforts can be facilitated if users have a template of what should be included in a test manual. Fortunately, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) outlines very clearly what data test authors and test publishers should provide for potential users. These *Standards* are fully described in Chapter 13 of this volume. At a minimum, test users need to know the characteristics of the norm group (e.g., numbers of children tested in each grade and their ethnicity and income status) and specific information that matches test items to content standards.

In contrast, other tests are low-stakes devices. Scores derived from such measures do not materially change the possibilities open to an individual or an organization. Cognitive screening devices fall into this category because scores from the screen are used only to determine additional testing, not access to particular services. Individuals identified by such tests as potentially needing services are reassessed before service delivery; thus, false positives (i.e., a

score that triggers services) are likely preferable to false negatives because all of the individuals who might need services should be identified in the screening model. The cost for identifying an individual who in the next round of testing is discovered not to need services (false positive) is small. The individual should not be harmed by the label of possibly needing services. However, if an individual is overlooked by a screening test (false negative), he or she would be harmed by not receiving enriched services. Therefore, the screening test is considered low stakes because the identification of all the people with a condition as well as a few more leads to more testing. The follow-up assessments become high stakes because of the decision to provide enriched services for the individual, a more precise undertaking.

The assessment processes involved in research are usually considered low-stakes assessments. For example, the development of assessments requires many individuals to take a test. Researchers seek documentation of the technical qualities of their assessment devices to investigate the internal and external validities of their test, but the test takers typically experience no repercussions as a result of a particular score.

CONCLUSION

The purpose of this chapter has been to contribute to the informed use of tests by alerting users to sources of information. A simple checklist of questions might be a useful summary for test users to consider as they identify a testing product. Before proceeding with the choice or use, each question must be considered.

1. What is the purpose or goal of the assessment?
2. What scores or variables are of interest?
3. Do tests advertised to meet the purpose and that derive the desired scores meet consensual standards for psychological and educational tests?
4. Is there related research that documents other aspects of score validity or reliability?
5. What do psychometric experts say about the tests in independent reviews?
6. Does the user have the training, time, and resources to administer and interpret the test with integrity?
7. Are the results of the tests likely to affect the test taker's material welfare? If so, special care is necessary.

Although many resources are available to assist test users in evaluating a test, the user retains the ultimate responsibility for deciding on the test's appropriateness for a specific application. Valid and ethical uses of tests require significant effort on the part of users. They must adopt attitudes of healthy skepticism toward advertisements from publishers. They must develop, or at least access, deep expertise about psychometric theory and research related to particular tests. Finally, they should make use of expert opinions and research related to particular instruments to ensure they act with the best available information as the basis for their ethical use of a test.

References

American College Testing Program. (1995). *Technical manual: Revised UNISEX edition of the ACT Interest Inventory (UNIACT)*. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Atkinson, R. C., & Schiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). San Diego, CA: Academic Press.

Becker, K. A. (2003). *History of the Stanford-Binet intelligence scales: Content and psychometrics* (Stanford-Binet Intelligence Scales, Fifth Edition Assessment Service Bulletin No. 1). Itasca, IL: Riverside.

Benton, A. L., & Hamsher, K. D., & Sivan, A. B. (1994). *Multilingual Aphasia Examination: Manual of instructions* (3rd ed.). Iowa City, IA: AJA Associates.

Betz, N. E., Borgen, F. H., & Harmon, L. W. (1996). *Skills Confidence Inventory applications and technical guide*. Palo Alto, CA: Consulting Psychologists Press.

Campbell, D. P., Hyne, S. A., & Nilsen, D. (1992). *Manual for the Campbell Interest and Skill Survey*. Minneapolis, MN: National Computer Systems.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Amsterdam, the Netherlands: Elsevier.

Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy, 9*, 882–888. doi:10.1016/S0005-7894(78)80020-3

Cramer, P. (1996). *Story-telling, narrative and the Thematic Apperception Test*. New York, NY: Guilford Press.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test*. San Antonio, TX: Psychological Corporation.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *The California Verbal Learning Test manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

DeStefano, L., & Thompson, D. S. (1990). Adaptive behavior: The construct and its measurement. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior, and context* (pp. 455–469). New York, NY: Guilford Press.

Exner, J. E., Jr. (2003). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations* (4th ed.). New York, NY: Wiley.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Mascolo, J. T. (2006). *The achievement test desk reference: A guide to learning disability identification* (2nd ed.). New York, NY: Wiley.

Franzen, M. D., & Iverson, G. L. (2000). The Wechsler Memory Scales. In G. Groth-Marnat (Ed.), *Neuropsychological assessment in clinical practice* (pp. 195–222). New York, NY: Wiley.

Gay, E. G., Weiss, D. J., Hendel, D. D., Dawis, R. V., & Lofquist, L. H. (1971). *Manual for the Minnesota Importance Questionnaire*. Minneapolis: University of Minnesota.

Glahn, D. C., Gur, R. C., Ragland, J. D., Censits, D. M., & Gur, R. E. (1997). Reliability, performance characteristics, construct validity, and an initial clinical application of a Visual Object Learning Test (VOLT). *Neuropsychology, 11*, 602–612. doi:10.1037/0894-4105.11.4.602

Golden, C. J., Purisch, A. D., & Hammeke, T. (1985). *Luria-Nebraska Neuropsychological Battery: Forms I and II*. Los Angeles, CA: Western Psychological Services.

Gronlund, N. E. (2003). *Assessment of student achievement*. Boston, MA: Allyn & Bacon.

Groth-Marnat, G. (Ed.). (2000). *Neuropsychological assessment in clinical practice: A guide to test interpretation and integration*. Hoboken, NJ: Wiley.

Groth-Marnat, G. (2009). *Handbook of psychological assessment*. Hoboken, NJ: Wiley.

Groth-Marnat, G., Gallagher, R. E., Hale, J. B., & Kaplan, E. (2000). The Wechsler Intelligence Scales. In G. Groth-Marnat (Ed.), *Neuropsychological assessment in clinical practice* (pp. 129–194). New York, NY: Wiley.

Hammill, D. D., Hresko, W. P., Ammer, J. J., Cronin, M. E., & Quinby, S. S. (1998). *Hammill Multiability Achievement Test (HAMAT)*. Austin, TX: Pro-Ed.

Hansen, J. C. (2005). Assessment of interests. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 281–304). New York, NY: Wiley.

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory applications and technical guide*. Palo Alto, CA: Consulting Psychologists Press.

Harrison, P., & Oakland, T. (2003). *Adaptive Behavior Assessment System* (2nd ed.). San Antonio, TX: Psychological Corporation.

Hebben, N., & Milberg, W. (2009). *Essentials of neuropsychological assessment* (2nd ed.). New York, NY: Wiley.

Helmes, E. (2000). Learning and memory. In G. Groth-Marnat (Ed.), *Neuropsychological assessment in clinical practice* (pp. 293–334). New York, NY: Wiley.

Holland, J. L. (1985). *Professional manual for the Self-Directed Search* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253–270. doi:10.1037/h0023816

Horn, J. L., & Cattell, R. B. (1982). Whimsy and misunderstandings of Gf–Gc theory. *Psychological Bulletin, 91*, 623–633. doi:10.1037/0033-2909.91.3.623

Hresko, W. P., Peak, P. K., Herron, S. R., & Bridges, D. L. (2000). *Young Children's Achievement Test*. Austin, TX: Pro-Ed.

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin, 136*, 422–449.

Jackson, D. N. (1984). *Manual for the Multidimensional Aptitude Battery*. Port Huron, MI: Research Psychologists Press.

Janowsky, J. S., & Thomas-Thrapp, L. J. (1993). Complex figure recall in the elderly: A deficit in memory or construction strategy? *Journal of Clinical and Experimental Neuropsychology, 15*, 159–169. doi:10.1080/01688639308402554

Johansson, C. B. (1986). *Career Assessment Inventory: The enhanced version*. Minneapolis, MN: National Computer Systems.

Johnstone, B., Holland, D., & Larimore, C. (2000). Language and academic abilities. In G. Groth-Marnat (Ed.), *Neuropsychological assessment in clinical practice* (pp. 335–354). New York, NY: Wiley.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement and practice* (pp. 127–167). Washington, DC: American Psychological Association. doi:10.1037/10063-004

Kaplan, E., Goodglass, H., & Weintraub, S. (1976). *Boston Naming Test: Experimental edition*. Boston, MA: Veteran's Administration Hospital.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children: Interpretive manual*. Circle Pines, MN: American Guidance Service.

Keyser, D. J., & Sweetland, R. C. (Eds.). (1984–2005). *Test critiques*. Austin, TX: Pro-Ed.

Kratochwill, T. R., Sheridan, S. M., Carlson, J., & Lasecki, K. L. (1999). Advances in behavioral assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 350–382). New York, NY: Wiley.

Kuder, G. F., & Zytowski, D. G. (1991). *Kuder Occupational Interest Survey, Form DD general manual*. Monterey, CA: California Testing Bureau.

Lezak, M. D., Loring, D. W., & Howieson, D. B. (2004). *Neuropsychological assessment*. Oxford, England: Oxford University Press.

Luckasson, R., Borthwick-Duffy, S., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., Reeve, A., . . . Tassé, M. J. (2002). *Mental retardation: Definition, classification, and systems of supports*. Washington, DC: American Association on Mental Retardation.

Marsh, H. W. (1994). Sport motivation orientations: Beware of jingle jangle fallacies. *Journal of Sport and Exercise Psychology, 16*, 365–380.

Morey, L. (1991). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Moye, J. (1997). Nonverbal memory assessment with designs: Construct validity and clinical utility. *Neuropsychology Review, 7*, 157–170. doi:10.1023/B:NERV.0000005907.34499.43

Nevill, D. D., & Super, D. E. (1986a) *The Salience Inventory: Theory, application, and research*. Palo Alto, CA: Consulting Psychologists Press.

Nevill, D. D., & Super, D. E. (1986b). *The Values Scale: Theory, application, and research*. Palo Alto, CA: Consulting Psychologists Press.

Nord, W. R., Brief, A. P., Atieh, J. M., & Doherty, E. M. (1990). Studying meanings of work: The case of work values. In A. P. Brief & W. R. Nord (Eds.), *Meanings of occupational work* (pp. 21–64). Lexington, MA: Lexington Books.

Ramsay, M. C., Reynolds, C. R., & Kamphaus, R. W. (2002). *Essentials of behavioral assessment*. New York, NY: Wiley.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 1. General overview*. San Antonio, TX: Harcourt Assessment.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Rey, A. (1964). *L'examen clinique en psychologie* [The clinical examination in psychology]. Paris, France: Presses Universitaires de France.

Roberts, R. D., Markham, P. M., Matthews, G., & Zeidner, M. (2005). Assessing intelligence: Past, present, and future. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 333–360). Thousand Oaks, CA: Sage. doi:10.4135/9781452233529.n19

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rounds, J. B., & Armstrong, P. I. (2005). Assessment of needs and values. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling* (pp. 305–329). New York, NY: Wiley.

Shapiro, E. S., & Kratochwill, T. R. (Eds.). (2000). *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed.). New York, NY: Guilford Press.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2006). *Vineland Adaptive Behavior Scales: Survey Interview Form* (2nd ed.). San Antonio, TX: Pearson Education.

Spearman, C. (1927). *The nature of "intelligence" and the principles of cognition* (2nd ed.). London, England: Macmillan.

Spies, R. A., Carlson, J. F., & Geisinger, K. F. (Eds.). (2010). *The 18th mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Trahan, D. E., & Larrabee, G. J. (1988). *Continuous Visual Memory Test manual*. Odessa, FL: Psychological Assessment Resources.

Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40*, 385–398. doi:10.1037/0003-066X.40.4.385

U.S. Department of Defense. (1984). *Test manual for the Armed Services Vocational Aptitude Battery* (DoD No. 1340.12AA). North Chicago, IL: U.S. Military Entrance Processing Command.

U.S. Department of Labor. (2000). *Work Importance Locator: User's guide*. Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor. (2002). *Work Importance Profiler: User's guide*. Washington, DC: U.S. Government Printing Office.

Watkins, C. E., Campbell, V. L., & Nieberding, R. (1994). The practice of vocational assessment by counseling psychologists. *Counseling Psychologist, 22*, 115–128. doi:10.1177/0011000094221008

Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology: Interdisciplinary and Applied, 19*, 87–95. doi:10.1080/00223980.1945.9917223

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised manual*. New York, NY: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002). *Wechsler Individual Achievement Test* (2nd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition: Technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8*, 231–258. doi:10.1177/073428299000800303

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psycho-Educational Battery—Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson Psychoeducational Battery* (3rd ed.). Chicago, IL: Riverside.

Zhu, J., Weiss, L. G., Prifitera, A., & Coalson, D. (2004). The Wechsler Intelligence Scales for Children and Adults. In G. Goldstein, S. R. Beers, & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment: Vol. 1. Intellectual and neuropsychological assessment* (pp. 51–75). Hoboken, NJ: Wiley.

CHAPTER 15

ETHICS IN PSYCHOLOGICAL TESTING AND ASSESSMENT

Frederick T. L. Leong, Yong Sue Park, and Mark M. Leach

Since their early origins in the use of intelligence tests for placement of schoolchildren through the recent attention to high-stakes educational testing, psychological testing and assessment have remained controversial and complex topics. This controversy underscores the importance of addressing the ethical challenges in the use and application of tests and assessment in psychology. In this chapter, we begin with an overview of the various professional ethical standards that guide our work in this area. This section is followed by a more detailed review and discussion of the relevant sections of the American Psychological Association (APA) *Ethical Principles of Psychologists and Code of Conduct* (APA, 2010). In this review, we also provide some guidance on the application of these ethical principles to the testing and assessment enterprise. Given the increasing cultural diversity of the U.S. population and the rise of globalization, we end with a discussion of some unique challenges in conducting testing and assessment cross-culturally.

There are also legal issues associated with testing and assessment in psychology, but these issues are not covered in this chapter because they are addressed elsewhere in this handbook (see Chapter 28, this volume, and Volume 2, Chapters 6 and 34). It is interesting to note that the U.S. Office for Human Research Protections highlights the differences between ethical principles and regulatory guidelines. *Ethical principles* refers to ethical values and principles aimed at the protection of human participants in research, whereas *regulatory guidelines* refers to a list of procedural dos and don'ts ("Distinguishing Statements of Ethical Principles and Regulatory Guidelines," 2011). The purpose of this chapter is to discuss the ethical values and principles in professional psychology as they pertain to testing and assessment.

PROFESSIONAL ETHICS

Ethics is a broad term that encompasses the commonly endorsed values of professional psychology (Groth-Marnat, 2006) and is the basis for ethics codes—rules and guidelines on appropriate behaviors for the purpose of protecting the public and the profession (Meara, Schmidt, & Day, 1996). In the United States, three major sources of ethics codes related to psychological testing and assessments are available: (a) the *Standards for Education and Psychological Testing* (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 1999), (b) the *Guidelines for Computer-Based Tests and Interpretations* (APA Committee on Professional Standards & Committee on Psychological Tests and Assessment, 1986), and (c) the *Ethical Principles of Psychologists and Code of Conduct* (APA, 2010).

Standards for Education and Psychological Testing

In 1985, AERA, APA, and NCME collaborated to develop the *Standards for Education and Psychological Testing*—a set of standards pertaining to professional and technical issues of test development and use in education, psychology, and employment. The

DOI: 10.1037/14047-015
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)

Copyright © 2013 by the American Psychological Association. All rights reserved.

Standards is organized in three sections: (a) Test Construction, Evaluation, and Documentation; (b) Fairness in Testing; and (c) Testing Applications. The *Standards* document was significantly revised in 1999 to contain a greater number of standards and updated to reflect changes in law and measurement trends, increased attention to diversity issues, and information on new tests and new uses of existing tests (AERA et al., 1999). An in-depth review of the *Standards* can be found in Chapter 13 of this volume.

Guidelines for Computer-Based Tests and Interpretations

With the increased use of, and concern for the lack of regulation of, psychological computer-based testing (CBT), APA's Committee on Professional Standards and Committee on Psychological Tests and Assessment (1986) published the *Guidelines for Computer-Based Tests and Interpretations*, a set of 31 guidelines aimed at both test developers, to ensure the development of quality CBT products, and end users of these products, to ensure proper administration and interpretation of computer-based psychological tests (Schoenfeldt, 1989). More recently, the International Test Commission gave increased attention to CBT in its own set of CBT guidelines, adopted in 2005, titled the *International Guidelines on Computer-Based and Internet-Delivered Testing*. Similar to the objectives of the *Guidelines for Computer-Based Tests and Interpretations*, the general aim of the International Test Commission guidelines is to recommend standards for good practices for development and use of CBTs. The International Test Commission guidelines are organized along the following recommendations: (a) Give due regard to the technological issues in computer-based and Internet testing, (b) attend to quality issues in CBT and Internet testing, (c) provide appropriate levels of control over CBT and Internet testing, and (d) make appropriate provision for security and safeguarding privacy in CBT and Internet testing.

American Psychological Association Ethical Principles of Psychologists and Code of Conduct

APA adopted its first official code of ethics in 1952 in response to the field's increased professionalism and visibility after World War II (Fisher, 2009). Since then, the APA Ethics Code has been revised 10 times, with an amended version being adopted in 2010 by the APA Council of Representatives. The APA Ethics Code contains four major sections. The first section, Introduction and Applicability, delineates the rationale, scope and limitations, and applicability of the Ethics Code and describes the possible consequences and sanctions imposed on APA members and student affiliates who are found to have violated the standards of the Ethics Code. The second section, the Preamble, contains a statement of APA's purpose as a profession and delineates the various roles and responsibilities held by psychologists. The third section, General Principles, contains the five aspirational general principles of APA meant "to guide and inspire psychologists toward the very highest ethical ideals of the profession" (APA, 2010, p. 3): Beneficence and Nonmaleficence, Fidelity and Responsibility, Integrity, Justice, and Respect for People's Rights and Dignity. Finally, the fourth section, Ethical Standards, contains a set of 10 enforceable ethical standards by which psychologists are obligated to abide. Sanctions may be imposed on psychologists who violate these ethical standards. The ninth section of the Ethical Standards provides guidelines pertaining to the use of psychological tests and assessments (APA, 2010). In the next section, we discuss the APA ethical standards on assessments in greater detail as they apply to a variety of purposes and contexts in which psychological testing is conducted.

AMERICAN PSYCHOLOGICAL ASSOCIATION ETHICAL STANDARDS ON ASSESSMENTS

In the sections that follow, we highlight the 11 assessment standards associated with the APA Ethics Code. These standards have been found in the ethics codes of other countries, although the degree to which there is consistency differs based on a country's use of testing. In addition, other countries did include an additional standard not found in the APA Ethics Code (Leach & Oakland, 2007). The consistency found indicates that these standards

have international appeal and form the ethical foundation of test use and development.

Bases of Assessments

APA Ethical Standard 9.01, Bases for Assessments, stipulates that all oral and written opinions and conclusions made by psychologists be based on information and techniques grounded in the scientific and professional knowledge bases of professional psychology (Fisher, 2009). Adherence to the scientific and professional standards of the field builds public trust in the profession consistent with Principle B, Fidelity and Responsibility, of the APA Ethics Code. When psychologists' opinions and conclusions are not grounded in the scientific and professional standards, the probability that their opinions may mislead and potentially harm the clients and patients whom they serve is greater. Professional discernment applies to all phases of the testing and assessment process, even in the preassessment phase of planning and information gathering (Jacob & Hartshorne, 2006).

Scientific and professional bases. According to APA Ethical Standard 9.01a, psychologists are obligated to base their recommendations, reports, and diagnostic or evaluative statements on techniques supported by the scientific and professional standards of the field. Moreover, Ethical Standard 9.01b stipulates that opinions on individuals' psychological characteristics be drawn after an adequate examination is conducted on the basis of assessment procedures and tools that are consistent with the objective of the testing (e.g., that address the referral question), are sensitive to the cultural and linguistic characteristics of the examinee, are congruent with the examinee's level of competency to be administered the assessment, and have been shown to be valid and reliable. Psychologists are responsible for personally ensuring that the reliability and validity of the assessment tools and techniques they use are adequate. Furthermore, psychologists should base their conclusions and recommendations on assessments that have been demonstrated to be reliable and valid. Reliability and validity issues are discussed in greater depth in Chapters 2 and 4 of this volume.

Limitations of assessment results. When limitations to the reliability and validity of the assessment procedures and tools are found, psychologists should appropriately limit the nature and extent of their conclusions and recommendations and refrain from drawing conclusions that are not adequately supported. Another scenario to limit conclusions may arise when psychologists are unable to personally evaluate an individual for various reasons, such as an examinee's refusal to continue with assessment or an examinee's relocation during the course of assessment. In these situations, psychologists should make reasonable efforts, when appropriate and practical, to reach examinees for assessment and thoroughly document the outcome of these efforts (Ethical Standard 9.01b). When a personal evaluation is not practical, psychologists are obligated to limit the scope of their decisions and recommendations, in addition to delineating how the limited information influences the reliability and validity of their findings.

Cases may exist in which personal evaluation of an examinee is not warranted, such as when reviewing preexisting records in academic, legal, organizational, and administrative contexts or when examining secondary records provided by a third-party assessor, such as trainees or professionals with whom psychologists supervise or consult, respectively (Fisher, 2009; Knapp & VandeCreek, 2003). In these cases, psychologists should clearly explain that their conclusions and recommendations are based on a secondary analysis of information derived from alternate sources (Ethical Standard 9.01c).

Use of Assessments

Psychological testing applies to a wide range of purposes and contexts, which include but are not limited to screening applicants for job placement, diagnosing psychological disorders for mental health treatment, verifying health insurance coverage, conducting focus groups for market research, informing legal decisions and governmental policies, and developing measures to reliably measure personality characteristics (Aiken & Groth-Marnat, 2006; Fisher, 2009). According to the *Eighteenth Mental Measurements Yearbook* (Spies, Carlson, &

Geisinger, 2010), there are no less than 19 major categories of psychological tests and assessments.

APA Ethical Standard 9.02 pertains to the proper selection and use of psychological tests and assessments. The first component of this ethical standard stipulates that psychologists administer, adapt, score, interpret, and use psychological testing in the manner and purpose for which the selected tests and assessments were designed to be used as indicated by research (Ethical Standard 9.02a). Furthermore, psychologists should select and use tests or assessments with members of populations for whom adequate reliability and validity of the test scores has been established. If the reliability and validity of the test scores has not been examined or verified for a particular population, psychologists are obligated to describe the strengths and limitations of the interpretations and recommendations derived from the test or assessment results (Ethical Standard 9.02b). The third aspect of this ethical standard obligates psychologists to select tests and assessments that are appropriate to the language preference and competence of the individuals being assessed (Ethical Standard 9.02c).

Test selection and usage. Psychologists are responsible for selecting appropriate assessments for the intended purpose of the testing (Ethical Standard 9.02a). To guide the selection of appropriate tests and assessments, psychologists should have adequate knowledge of the theoretical bases and empirical evidence that support the validity and reliability of the tests or assessments; standardized administration and scoring procedures; approaches to interpreting the results; and the populations for which the assessment was normed and designed (Fisher, 2009; see Ethical Standard 9.07, Assessment by Unqualified Persons). Psychologists should also keep themselves updated on the most recent versions of the tests and assessments that they commonly use because testing and assessment procedures and parameters may change in light of theoretical advances and new research (see Ethical Standard 9.08, Obsolete Tests and Outdated Test Results). Finally, psychologists should select tests and assessments that have been empirically validated to be used in the specific contexts and settings in which the testing occurs.

Testing across diverse populations. According to Principles D (Justice) and E (Respect for People's Rights and Dignity) of the APA Ethics Code, psychologists strive to establish fair and equal access to and benefit of psychological contributions for all individuals and populations, which include but are not limited to diversity in age, gender, gender identity, race, ethnicity, culture, national origin, religion, disability, language, and socioeconomic status. Although psychological testing represents a unique contribution of professional psychology to benefiting larger society, ensuring the fair and equal access to and benefit of psychological testing has historically been challenging for the field. According to Reynolds (1982), the reliability and validity of test and assessment scores have predominately been established with White, middle-class samples and may not generalize well to other populations, especially those that represent a minority in the United States. This historical precedence conflicts with Ethical Standard 9.02b, which stipulates the selection and use of assessments that have been found to be adequately valid and reliable for drawing particular inferences for specific populations being assessed. When tests are administered across diverse populations, psychologists are obligated to select and use tests and assessments that have measurement equivalence in that the psychometric properties (i.e., measurement and structural models) have been shown to be equivalent or invariant between members of culturally different populations and those from the reference population for which the test and assessment scores were validated, normed, and found to be reliable (Schmitt, Golubovich, & Leong, 2010).

Testing and language. APA Ethical Standard 9.02c stipulates that psychologists select tests that are appropriate to be used with the language preferences and levels of competence of the individuals or groups being assessed. Thus, before selecting assessments, it is helpful for psychologists to gather information on examinees' cultural background (e.g., acculturation) and native and English language ability with regard to written, reading, and spoken language proficiencies (Jacob & Hartshorne, 2006; Takushi & Uomoto, 2001). According to Groth-Marnat (2009), literal translation of testing

and assessment materials and tools using the commonly implemented method of translation–back-translation may not be adequate because of cross-cultural differences in the conceptual interpretation of items, noncomparable idioms, and within-group differences in dialect and word usage. Furthermore, from an item response theory framework, literal translation of testing and assessment items from one language to another may change the properties of the items' difficulty, which may in turn diminish the measurement equivalence of tests or assessments. For these reasons, the psychometric properties of the original-language version of tests or assessments cannot be assumed to generalize to the alternate-language versions that were developed from a translation–back-translation method. More information on testing and language can be found in Volume 3, Chapter 26, of this handbook.

With regard to testing conducted in person (e.g., interviews) with linguistically different clients, psychologists may consider enlisting the services of a translator for interpretation purposes or consider referring clients to colleagues who have professional proficiency in the clients' language. Professional organizations may be useful resources for identifying and referring clients to professional colleagues with the appropriate linguistic background; for example, the National Association of School Psychologists maintains a directory of bilingual school psychologists that can be found on its website (http://www.nasponline.org/about_nasp/bilingualdirectory.aspx).

Informed Consent in Assessments

Before administering an assessment, psychologists are obligated to obtain from examinees, or their parents, guardians, or legal representatives, informed consent that includes an explanation of the nature and purpose of the assessment, fees, involvement of third parties (e.g., referral source), and limits of confidentiality (see Ethical Standard 3.10, Informed Consent). The informed consent stage of testing may also be the opportune time to provide examinees with an explanation of their rights as test takers. The Joint Committee on Testing Practices (1998) developed the *Rights and Responsibilities of Test Takers: Guidelines and Expectations* to inform test takers about and clarify expectations for the testing process. Because *consent* refers to examinees' legal status to autonomously decide whether to be assessed, informed consent must be communicated in a clear and comprehensible manner that is appropriate to the age of examinees and their mental abilities (Fisher, 2009).

As stipulated by Ethical Standard 9.03a, informed consent can be dispensed with in the following situations: when "(1) testing is mandated by law or governmental regulations; (2) informed consent is implied because testing is conducted as a routine educational, institutional or organizational activity; or (3) one purpose of the testing is to evaluate decisional capacity" (APA, 2010, p. 12). Even though informed consent is not required in these cases, psychologists are recommended to, when appropriate, continue to provide examinees with an explanation of the nature and purpose of the testing.

When assessing individuals younger than age 18 (i.e., minors), informed consent from parents or legal guardians is required because minors are viewed, from a legal standpoint, as being unable to make autonomous and well-informed decisions pertaining to psychological services. Thus, minors do not have the legal right to assent, consent, or object to a proposed psychoeducational assessment; however, it is recommended that minors be fully informed about the nature and purpose of the testing and assessment in a clear and understandable manner (Jacob & Hartshorne, 2006).

Nature and purpose of assessment. Informed consent in the assessment context includes an explanation of the nature and purpose of the test or assessment. Thus, psychologists are obligated to clearly explain how results will be used, the administration procedure, and possible benefits and risks or consequences of being assessed. With regard to informing examinees about the administration procedure, psychologists are advised to provide a general description of the procedure because foreknowledge of the testing may influence examinees' responses and thus alter the validity of the test or assessment results. Psychologists should also be sensitive to the possible risks and consequences of the testing, especially with regard to the negative

feelings that may be generated by the testing process. Some assessment topics or questions may elicit uncomfortable feelings in examinees, such as those that involve private or taboo topics (Groth-Marnat, 2009). Thus, psychologists, in most cases, should not pressure or force examinees to answer all questions, especially those that create undue discomfort or emotionally painful feelings.

Confidentiality and release of information. A core component of informed consent is explaining the limits of confidentiality. *Confidentiality* refers to a professional standard that requires psychologists to maintain the privacy of any assessment information unless disclosure is permitted or requested by examinees through a release of information. According to Ethical Standard 4.05, Disclosures, psychologists may breach confidentiality without examinees' permission when disclosure is mandated by law or when permitted by law for a valid purpose, such as to

> (1) provide needed professional services; (2) obtain appropriate professional consultations; (3) protect the client/patient, psychologist, or others from harm [e.g., danger to self and others, elder and child abuse]; and (4) obtain payment for services from a client/patient, in which instance disclosure is limited to only information that is necessary to obtaining the payment. (APA, 2010, p. 7)

In situations in which breach of confidentiality is necessary or legally mandated, psychologists should share only information that is necessary to accomplish the purpose of the disclosure in an effort to respect examinees' right to privacy.

Health Insurance Portability and Accountability Act and Family Educational Rights and Privacy Act. Because of the increased reliance on electronic databases to store client–patient information, psychologists are responsible for effectively protecting the confidentiality and security of the information contained in these databases (Aiken & Groth-Marnat, 2006). The Health Insurance Portability and Accountability Act (HIPAA) was established in 1996 to regulate the protection of protected health information. *Protected health information* refers to any information that

> (a) is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (b) relates to the past, present, or future physical or mental health or condition of any individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual. (Title 42, U.S.C. § 1320d)

Any health care provider who electronically transmits health information is considered a covered entity by HIPAA and must comply with HIPAA regulations. Within the informed consent, covered entities should provide examinees with a written document titled *Notice of Practice Practices;* this document contains a description of the examinee's rights, the legal duty to protect protected health information, and the routine uses and disclosures of protected health information. The Family Educational Rights and Privacy Act of 1974 pertains to issues of confidentiality and release of information in the educational setting. The act stipulates that assessment information and school records of students maintained by educational institutions that receive federal funding may be disclosed to others only with the written consent of the student examinees or their parents or legal guardians.

Language and use of interpretation services. Ethical Standards 9.03b and 9.03c refer to the psychologists' responsibility to provide informed consent in the language of the examinee or at a language proficiency level the examinee can reasonably understand. Psychologists may enlist the services of an interpreter when working with examinees who have limited English proficiency. When using interpreters, psychologists are responsible for ensuring that interpreters are not only competent in communicating the informed consent in a reasonable and understandable manner but also comply with the ethical standard on maintaining the confidentiality of examinees' identity, assessment results, and

test security (Fisher, 2009; Knapp & VandeCreek, 2003).

Release of Test Data

According to Fisher (2009), a growing trend in the legal system is toward affirming the autonomy of patients' access to their health care records, a trend that is consistent with Principle E, Respect for People's Rights and Dignity, of the APA Ethics Code, emphasizing self-determination. HIPAA stipulates that patients have the right to access, inspect, and receive copies of their medical and billing records on their request for the release of this information. Related to the assessment context, examinees or others identified in the release have the right, in most cases, to have access to their test data (Ethical Standard 9.04a). *Test data* refers to raw and scaled scores on the assessment items, any responses to test questions or stimuli, and psychologists' written notes or recordings of the testing.

Test data versus test materials. It is important to note the difference between test data and test materials. *Test materials* refers to test manuals, administration and scoring protocols, and test items. According to Ethical Standard 9.11, test materials do not need to be released pursuant to a client or patient request for test data because test materials are protected by copyright laws, and inappropriate release of such test materials is legally considered a breach of trade secrets (Groth-Marnat, 2009; Knapp & VandeCreek, 2003). However, when examinees' identifying information or responses are written on test materials, the test material is considered test data and may need to be released on examinees' request (Ethical Standard 9.04a). Thus, examiners are recommended, whenever possible, to record any identifying information and responses on a separate document from the actual test materials.

Potential misuse of test data. When examinees provide a release to request test data for themselves or identified others, it is important that psychologists explain the potential for test data to be misused if the people interpreting the test data do not have the proper qualifications to do so (see Ethical Standard 9.07, Assessment by Unqualified Persons). According to Ethical Standard 9.04a, psychologists may refrain from releasing test data to the examinees or others if the release may result in substantial harm resulting from misuse or misinterpretation of the test data. In these cases, psychologists are obligated to document the specific rationale for why they believe that the test data would result in substantial harm (Fisher, 2009).

Court order for test data. According to Ethical Standard 9.04b, psychologists are obligated to release test data when the disclosure is required by the law or court order. When release of test data is court mandated, Fisher (2009) recommended that psychologists seek legal counsel to determine the legitimacy of the request and ascertain their legal responsibility to release the test data. Another recommendation is that psychologists request the court for a protective order to prevent the inappropriate disclosure of the confidential test data and recommend that test data be reviewed by another health care professional who is qualified to provide appropriate and competent interpretations. Furthermore, psychologists are recommended to make reasonable efforts to notify examinees when test data are released to the court and to document these efforts (Fisher, 2009).

Test Construction

Ethical Standard 9.05, Test Construction, refers to test developers' responsibility to ensure that the development of tests and assessments incorporates appropriate psychometric procedures that are guided by the current scientific and professional knowledge of test design, standardization, validation, reduction or elimination of bias, and recommendations for use.

Standardization. Test developers are responsible for providing specific and clear guidelines to qualified test users with regard to the proper and standardized procedure for administering and scoring tests and assessments. Furthermore, test developers are responsible for specifying the scoring cutoffs and norms for the populations for which the tests and assessments were developed and intended to be used. Scoring norms are commonly found in norm-referenced tests, which allows for comparison of individual scores to the distribution of scores from

the reference group. It is important that the characteristics of the reference group sample are clearly described in the test or assessment manual and are representative of the population to which the test is targeted.

Validity. According to the *Standards for Educational and Psychological Tests* (AERA et al., 1999), *validity* is defined as the degree to which the theoretical basis for the assessment and accumulated empirical evidence support the intended interpretation of the scores for which the assessment was designed. In general, validity refers to the degree to which an assessment measures what it purports to measure. Several types of evidence are used to justify claims of validity, such as content-related evidence and criterion-related evidence. For an in-depth review, readers are referred to Chapter 4 in this volume.

Reliability. The *Standards for Educational and Psychological Tests* (AERA et al., 1999) stipulate that test developers are obligated to provide reliability estimates—the degree to which the assessment results are consistent over repeated administrations—of their tests and assessments. Jacob and Hartshorne (2006) recommended that reliability estimates be provided for each demographic subpopulation of the population for which the assessment was intended, such as for age groups and class levels. Several methods can establish the reliability of an assessment: internal consistency, test–retest, split-half test, and alternative-form comparisons. For an in-depth review, readers are referred to Chapter 2 in this volume.

Interpreting Assessment Results

Interpretations of test and assessment results influence the decisions and recommendations that are made in reference to the purpose of the testing (see Ethical Standard 9.02, Use of Assessments), such as diagnosing and informing treatment plans in clinical settings and educational placements in academic settings and determining employment selections and promotions. Interpretations should be based on proper administration of tests and assessments as outlined by the testing manual to ensure the interpretations are in line with the evidence to support the validity and reliability of the test or assessment scores (Fisher, 2009). It is the psychologist's responsibility to ensure that his or her interpretations of test or assessment results are useful and relevant to the purpose of the assessment and take into account various test factors, test-taking abilities, and other characteristics of individuals being assessed (Ethical Standard 9.06).

Interpretation of multiple sources. Interpretations of test and assessment results should not be derived from a simple, mechanical process that is based solely on the test or assessment scores, score cutoffs, or reliance on automated interpretations (Fisher, 2009; Groth-Marnat, 2009) but that takes into consideration a host of factors, including but not limited to examinees' characteristics, test-taking abilities, styles, issues of fatigue, perceptual and motor impairments, illnesses, language proficiencies, and cultural orientations (Fisher, 2009). Furthermore, Groth-Marnat (2009) recommended that psychologists base their interpretations on multiple sources of data, including behavioral observations, examinee background information, and other assessments. Often, testing is administered using an integrated battery of assessments, and inconsistent findings across the various assessments may result. In these situations, it is the psychologist's responsibility to analyze the contradictions and use his or her clinical and professional judgment to offer the most accurate and relevant interpretation in relation to the purpose of testing (Groth-Marnat, 2009).

Automated interpretations. There are many well-established, standardized assessments, such as the Minnesota Multiphasic Personality Inventory—2, for which one can receive a computer-generated automated interpretative report. Although these automated interpretations are based on a body of past empirical evidence and theoretical models, it is important to highlight that interpretations are not sophisticated enough to take into account examinees' unique characteristics and test-taking contexts. Thus, psychologists should not base their interpretations solely on automated interpretations but rather use automated interpretations as supplemental resources for integrated interpretations that take into consideration a host of other factors that may influence the testing.

Limitations of interpretations. According to Ethical Standard 9.06, Interpreting Assessment Results, psychologists are obligated to indicate any significant limitations of their interpretations, especially when the interpretations are not supported by the established validity and reliability of the test or assessment scores in making particular inferences. When interpretation of test or assessment scores is made outside their established validity and reliability, Fisher (2009) recommended that such interpretations be posed as hypotheses, rather than conclusions, to elucidate the limitations of such findings. Another limitation that needs to be indicated is when testing procedures and materials, evidence for validity and reliability, and score cutoffs and norms have become obsolete in the face of new research or changes in the populations for which tests and assessments were designed (see Ethical Standard 9.08, Obsolete Tests and Outdated Test Results).

Assessment by Unqualified People

APA Ethical Standard 9.07, Assessment by Unqualified Persons, warns against the promotion of psychological assessment techniques being used by unqualified people. Psychologists are obligated to ensure that testing is carried out by qualified individuals within the scope of their competence as indicated by their education and training background and past experiences (Fisher, 2009). Furthermore, qualified psychologists have knowledge of the nature and purpose of the assessments, their psychometric properties, standardized procedure for administration and scoring, proper interpretation of results, and assessment limitations (Groth-Marnat, 2009). Unqualified users may also include psychologists who are working with populations or problem areas that are outside the scope of their competencies (see Ethical Standard 2.01, Boundaries of Competence), such as working with culturally and linguistically different clients whom they are not multiculturally competent to serve.

Assessment by unqualified people may result in misdiagnosis of the examinees' presenting concerns and potentially result in psychological harm (Jacob & Hartshorne, 2006). Aiken and Groth-Marnat (2006) suggested that the unqualified use of assessments has greater consequences when assessing individuals (e.g., intelligence and personality assessments) as opposed to groups because misuse of assessment results can have direct negative consequences on people's livelihoods, such as being prescribed a treatment plan for an incorrect diagnosis or being placed at the wrong educational level or in the wrong job placement. In relation to Principle A, Beneficence and Nonmaleficence, of the APA Ethics Code, psychologists should be aware of the boundaries or limitations of their competence to prevent unqualified use of assessments and make appropriate referrals or seek supervision or consultation from specialists in these situations (Aiken & Groth-Marnat, 2006). Furthermore, psychologists are recommended to obtain access to or create a directory of local assessment specialists for referral purposes (Jacob & Hartsthorne, 2006).

Qualifications. According to Turner, DeMers, Fox, and Reed (2001), qualified use of assessments often includes graduate course work and supervised training experiences pertaining to the use of specific assessments. In 2002, the Psychological Assessment Work Group convened at the Competencies Conference: Future Directions in Education and Credentialing in Professional Psychology and identified a set of eight core competencies in psychological testing:

1. A background in the basics of psychometric theory.
2. Knowledge of the scientific, theoretical, empirical, and contextual bases of psychological assessment.
3. Knowledge, skill, and techniques to assess the cognitive, affective, behavioral, and personality dimensions of human experience with reference to individuals and systems.
4. The ability to assess outcomes of treatment/intervention.
5. The ability to evaluate critically the multiple roles, contexts, and relationships within which clients and psychologists function, and the reciprocal impact of these roles, contexts, and relationships on assessment activity.
6. The ability to establish, maintain, and understand the collaborative professional relationship that provides a context for all psychological activity including psychological assessment.

7. An understanding of the relationship between assessment and intervention, assessment as an intervention, and intervention planning.
8. Technical assessment skills that include: (a) problem and/or goal identification and case conceptualization, (b) understanding and selection of appropriate assessment methods including both test and non-test data (e.g., suitable strategies, tools, measures, time lines, and targets), (c) effective application of the assessment procedures with clients and the various systems in which they function, (d) systematic data gathering, (e) integration of information, inference, and analysis, (f) communication of findings and development of recommendations to address problems and goals, (g) provision of feedback that is understandable, useful, and responsive to the client, regardless of whether the client is an individual, group, organization or referral source. (Krishnamurthy et al., 2004, pp. 732–733)

The Psychological Assessment Workgroup also delineated core competencies of training programs in providing quality educational and training experiences for psychological testing.

Ethical responsibility for qualified use applies not only to individual psychologists but also to test developers with regard to the distribution of their test materials. Standards for qualified use have been established by test developers to prohibit unqualified users' access to test materials. Thus, test developers should include information on the required qualifications for use in the test's promotional materials and require end users to meet the minimum requirements to purchase and use their tests and assessments. Aiken and Groth-Marnat (2006) provided a sample qualification form for test developers that includes questions for the potential end user with regard to the purpose for using the test, area of professional expertise, level of training, specific courses taken, and quality control over test use (e.g., test security, appropriate tailoring of interpretations).

Assessment by trainees. Although APA Ethical Standard 9.07 stipulates that psychologists should not promote unqualified use of assessments, an exception is made for training purposes as long as trainees have adequate supervision while the assessments are provided. More specifically, for trainees to be qualified in administering tests or assessments, they must have been or concurrently be enrolled in a graduate-level course, practicum externship, or pre- or postdoctoral training program that provides training in the specific assessment that is being administered. In addition to the formal training, trainees must receive adequate supervision from a qualified user of the test or assessment. In cases in which unqualified trainees have not received sufficient training and supervision to administer the assessment, they must clearly inform examinees that the test or assessment is being administered for training purposes only and adequately describe the limitations of their assessment interpretations, conclusions, and recommendations (Fisher, 2009). It is important to note that when supervising psychologists sign their trainees' assessment reports, they are ultimately held responsible for the contents of the report (Jacob & Hartshorne, 2006).

Obsolete Tests and Outdated Test Results

Psychologists are prohibited from basing their decisions and recommendations on test data that are outdated for the test's current use (Ethical Standard 9.08a) and from tests and assessments that are obsolete and not useful for the current use (Ethical Standard 9.08b). Use of outdated test data is prohibited because examinees may have changed since the time of the prior assessment owing to such factors as maturational and developmental effects, development of new presenting problems, and changes in the environment (Fisher, 2009). When outdated test results are used, psychologists are obligated to provide an explanation for why outdated test data are used and to clearly communicate the limitations of such outdated information.

Old test data are often kept stored in outdated files or databases even after examiners no longer work at the testing location. In this situation, psychologists are recommended to prevent the misuse of outdated test results by taking reasonable steps to remove or destroy obsolete data and files. In cases in which clients or patients request that outdated test

data be sent to a new clinician who is currently providing services to them, psychologists are recommended to include a cover page detailing the limitations of outdated test results.

APA Ethical Standard 9.08 also stipulates that psychologists should not base their decisions and recommendations on use of obsolete assessments. According to Fisher (2009), tests developers often revise their assessments to reflect significant advances and changes in the theoretical constructs underlying the psychological characteristics being assessed; changes in the assessment's test item validity owing to various cultural, educational, linguistic, or societal influences; and shifts in the demographics of the target population, which in turn affect the standardized norms and score cutoffs. Use of obsolete tests may be applicable when long-term comparisons of test performance are needed, but psychologists are obligated to adequately describe the differences between test versions and explain the limitations of their comparisons when obsolete tests are used. According to Fisher (2009), the expense associated with updating to new versions is not an adequate ethical justification for using obsolete tests and assessments.

Test Scoring and Interpretation Services

APA Ethical Standard 9.09 applies to psychologists who provide test scoring and interpretation services. Within their promotional and other administrative materials (e.g., manuals), these psychologists are obligated to accurately describe the nature and purpose of the assessments, the basis for the standardized norms, and validity and reliability information for their assessment results and interpretations and to specify the qualifications for using the services. When interpretations and recommendations from assessment results are made, psychologists are obligated to provide the theoretical rationale and psychometric evidence for justifying their conclusions and to adequately explain the limitations of their interpretations and recommendations.

Ethical responsibility for the appropriate use of test scoring and interpretation services also applies to psychologists who are consumers of these services. These psychologists are obligated to select services that adequately provide evidence for the validity and reliability of their procedures for administering, scoring, and interpreting test and assessment results. Furthermore, psychologists using these services are obligated to have the qualifications and competence to ensure that the scoring and interpretations made by these services are consistent with APA Ethical Standard 9.06, Interpreting Assessment Results. When these services are used, the HIPAA Notice of Privacy Practices obligates psychologists to inform and obtain authorization from their clients or patients to permit the release of test or assessment information to these services.

Explaining Assessment Results

According to Ethical Standard 9.10, Explaining Assessment Results, psychologists are obligated to provide competent feedback to examinees, or to parents or legal guardians of minors, explaining any interpretations, decisions, and recommendations in relation to the purpose of testing. Groth-Marnat (2009) recommended that the feedback begin with a clear explanation of the rationale for testing, followed by the nature and purpose of the assessment, general conclusions drawn from assessment results, limitations, and common misconceptions or misinterpretations of assessment results. When examinees are minors, psychologists are obligated to provide the feedback to both examinees and their parents or legal guardians.

Sensitivity in the communication of assessment results. The *Standards for Educational and Psychological Tests* (AERA et al., 1999) stipulates that simple, clear, everyday language should be used when providing feedback so that the feedback is readily understood by its recipients. Psychologists should tailor their level of communication to recipients' personal characteristics, such as their educational and linguistic backgrounds, level of knowledge of psychological testing, and possible emotional reactions to the assessment results (Groth-Marnat, 2009). With regard to the possible emotional reactions generated by feedback, it may be helpful for psychologists to make available options for follow-up counseling to facilitate services for examinees who may need support in processing the feedback information. When providing

feedback on mental health status, Aiken and Groth-Marnat (2006) recommended that the least stigmatizing label be used to describe the examinees' psychological conditions or diagnoses.

Written reports. In addition to the oral feedback session, psychologists commonly provide written reports to examinees, or their referral source, regarding the assessment results, interpretations, and recommendations. Written reports should be centered on referral questions and the purpose of the testing and adequately describe the characteristics of the examinees and how they relate to the assessments used and the test situations (Aiken & Groth-Marnat, 2006). According to Jacob and Hartshorne (2006), written reports should be comprehensible to both professionals and nonprofessionals and should be written in a succinct, clear, and comprehensible manner while avoiding overgeneralizations (Aiken & Groth-Marnat, 2006). Psychologists are responsible for signing off on assessment reports only after ensuring the accuracy of the contents contained in the reports.

Maintaining Test Security

According to Ethical Standard 9.11, Maintaining Test Security, psychologists are obligated to maintain the security of test materials, which are defined as manuals, instruments, protocols, and test questions or stimuli. As noted in Ethical Standard 9.04, although examinees have the right to request and access test data, they do not have the right to access test materials for reasons related to threats to validity and copyright protection. For these reasons, test materials should be stored in a secure location, and only authorized and qualified individuals should have access to them. Furthermore, test materials, even sample items, should not be reprinted in any form, such as in newspapers and magazines, without the written consent of the test developers.

Threat to validity. A primary reason for the ethical obligation to maintain test security is the threat to test validity that is posed when individuals have access to test materials before administration of the test. Having foreknowledge of the test questions and answers may alter the psychometric properties of the test, including its standardized score cutoffs and norms and validity (Fisher, 2009). Furthermore, access to test materials before administration may increase the likelihood of some individuals manipulating their responses for purposes of malingering or obtaining an unfair advantage on a given assessment relative to others (Knapp & VandeCreek, 2003).

Copyright law. Pursuant to copyright protection laws, it is illegal and an ethical violation to reproduce test materials without obtaining permission from test developers or publishers. Maintaining test security allows for the protection of trade secrets and honors the terms of agreement made with the test publisher on obtaining access to the test materials (Groth-Marnat, 2009). With regard to HIPAA, which stipulates that examinees have the right to access their protected health information (e.g., test data), psychologists should separate, when appropriate, test materials from test data to protect the copyrighted test materials from being disclosed when releases of information are requested by clients or patients.

CROSS-CULTURAL ISSUES

Testing and assessment become inherently more complex when considering cross-cultural issues. Our position is that to be ethically and multiculturally competent when conducting testing and assessments, the psychologist should consider the client's cultural context. Approximately one third of the U.S. population consists of ethnic minorities, and when one includes the potential influence of other diverse groups (e.g., language, disability, socioeconomic status) on testing, it becomes evident that to be competent in testing and assessment requires much more than basic knowledge of test use.

All of the principles described in APA's (2010) *Ethical Principles of Psychologists and Code of Conduct* apply to cross-cultural testing, yet two are briefly highlighted that seem particularly salient. These are Principle D (Justice) and Principle E (Respect for People's Rights and Dignity). First, Principle D refers not only to equal access and fairness but to psychologists' ensuring that their biases, boundaries of competence, and level of expertise do

not influence their work and lead to unjust practices. Second, Principle E refers to respecting differences among individuals and cultural groups and the belief in autonomous self-determination. Unfortunately, sound ethical testing practices have not always been the norm when considering the history of the testing movement in psychology. Although progress in ethical testing practices has been made over the years and the field has improved significantly in the development, measurement, and implementation of testing with regard to culture, further developments are needed.

Psychological testing has made great strides in the understanding of psychological constructs, and it continues to do so. It also has a well-referenced history of bias against those who are not White, middle class, and male. The acceptance of the belief in universality, that the mainstream American experience is applicable to everyone, has long been at odds with a multicultural framework. This framework states that testing and assessment cannot be uniformly applied to all groups (Leong, Qin, & Huang, 2008). Using a simple example, readers would probably agree that assessing women if a test was normed on men or adults if a test was normed on elementary school-aged children would not be ethically appropriate. Similarly, there may be concerns about the application of tests primarily normed on the dominant group when considering use with nondominant group members. Consistent with many psychologists today, Burlew (2003) cautioned against taking a universal philosophical approach in that theories may not be transferable across cultures, that researchers are limited from developing alternative theories, that protective measures unique to a particular cultural group are neglected, and that any deviation from the universal perspective leads to a pathological or deviational view of nondominant outgroups. Only during the past few decades has research attention been given to the inclusion of diverse individuals and groups as they relate to the richness in understanding human behavior.

Etic Versus Emic

Validity from a cross-cultural perspective begins with knowledge of differences between etic and emic approaches to testing. Simply defined, etic approaches assess constructs across cultures, whereas emic approaches examine a construct within a particular culture. Understanding these validity issues is crucial when developing or using tests because tests are generally developed within a particular cultural context. Both etic and emic approaches are discussed in greater detail next, and examples from history are included to highlight ethical issues that have emerged.

Etic

Psychological testing has been at the forefront of controversy since the early part of the 20th century because of differences found among ethnic groups on a variety of tests, most notably intelligence tests. Imposed etics surrounding psychological assessment probably began with Galton's (1883/2003) treatise, "Inquiries Into Human Faculty and Its Development." This document led to the "mental test," which then helped launch psychology's version of the eugenics movement (Schultz & Schultz, 2011). Other psychologists such as Cattell, Goddard, and Terman were influential in launching intelligence and ability testing into conventional psychology. These famous psychologists, along with other equally as recognizable names such as Yerkes, were influential in putting forth testing practices that were unfavorable toward ethnic minorities, those of lower socioeconomic status, and others. More recently, Herrnstein and Murray's (1994) controversial book *The Bell Curve* revived the debate over the relationship among (primarily ethnic) groups and intelligence. Their thesis that ethnic minorities do not score well on tests of intelligence and achievement because of genetic and biological limitations harkens back to earlier testing history in psychology (for a review of the issues surrounding *The Bell Curve* and a rebuttal, see Jacoby & Glauberman, 1995).

Culturally appropriate and ethical test development has recently gained significant attention in the professional literature (e.g., Dana, 2005; Groth-Marnat, 2009). In this vein, to work toward competent, ethical, and culturally valid testing practices, psychologists and others have begun discussing test equivalence (or invariance). *Equivalence* refers to the

degree to which the parameters of a test's measurement model are comparable across groups (Cheung, van de Vijver, & Leong, 2011). Measurement equivalence is a prerequisite before one can make reasonable and ethical interpretations of the results across cultural groups. Historically, equivalence in psychological testing was omitted or significantly flawed given that many psychological tests were either normed on or developed in a framework of the dominant culture. Quite simply, using a psychological test that has not included a broader multicultural framework may introduce bias and is ethically dubious. It may be unethical because, among a myriad reasons, the psychologist is not acting competently and the foundation on which the tests were developed is flawed. More specifically, the APA Ethics Code acknowledges that ethical test use requires that the test be appropriate for the individual or group under investigation. Determination of whether a psychological instrument is valid for use with a particular cultural group is based on multiple factors, such as an individual's level of acculturation, translation of the instrument, language abilities, whether the construct measured with the instrument is consistent across cultures, and norm availability, among others. These can be accomplished through the assessment of four types of equivalence: linguistic, conceptual, metric, and functional (Leong, Leung, & Cheung, 2010).

Linguistic Equivalence

Linguistic equivalence, or translation equivalence, is primarily concerned with the translation of a psychological instrument and its application in another culture (Groth-Marnat, 2009). Brislin (1970) was one of the first to discuss the back-translation method, which involves translating an instrument into another language and then back-translating it into the primary language. The two versions are compared, and differences are resolved. Linguistic equivalence merely permits comprehensibility and does not, however, postulate about the instrument's validity. It is still a common translation method, although more recent procedures regarding the area of linguistic equivalence are expounded on in Hambleton, Merenda, and Spielberger (2005) and Volume 3, Chapter 26, of this handbook.

Conceptual Equivalence

Unfortunately, linguistic equivalence may be sufficient with some tests, but conceptual equivalence is also needed to behave in the highest ethical manner. Conceptual equivalence determines the degree to which a concept is consistent cross-culturally. This concept is more difficult to attain because what may be considered a similar concept between cultures may actually be a close proximity to it or interpreted differently altogether, resulting in conceptual variability. To decrease this variability, Usunier (1998) suggested that the translation process include multiple sources and target languages. Briefly, multiple native speakers independently develop words consistent with a concept, and a cross-cultural research team identifies the most commonly cited terms and back-translates them. Etic and emic conceptual dimensions are then determined (see also Leong et al., 2010).

Metric Equivalence

Metric equivalence is concerned with whether the psychometric properties of an instrument are consistent across cultural groups (Groth-Marnat, 2009). This type of equivalence is delineated into two categories, measurement invariance and structural invariance. Measurement invariance is related to variables' relationships to latent constructs, whereas structural invariance involves the actual latent variables themselves. Another way of considering the two is that measurement invariance is concerned with consistent matrices and scalar equivalence, for example, whereas structural invariance is concerned with whether the structural models, for example, are consistent across cultural groups. The more metric variability introduced, the greater the likelihood is that using the test across cultures is invalid and unethical.

Functional Equivalence

Functional equivalence addresses the idea that patterns of relationships between various constructs and a target measure are equivalent. If one construct in one culture does not function in the same manner in another culture, then variability is increased. For example, cognitive distortions may be associated with depression in one culture but not in another.

To test for cognitive distortions in one culture because of its cultural consideration as a common feature of depression in another culture could be inaccurate. To derive meaning and make interpretations from test results based on functional invariance could be considered unethical behavior (for a brief overview of strategies to offset measurement inequivalence, see Leong et al., 2008, 2010).

At least five ethical standards should be considered when evaluating tests without equivalence. We first consider a translated test developed in the English language and administered, for example, to an individual whose native language is Spanish. As indicated earlier, Ethical Standards 9.01, 9.02, and 9.06 are directly related to test use, and these three standards are central to linguistic equivalence. Standard 9.01, Bases for Assessments, states, "Psychologists base the opinions contained in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, on information and techniques sufficient to substantiate their findings" (APA, 2010, p. 12). Without linguistic equivalence, for example, a simple translation without the back-translation, the psychologist is acting unethically because whether the translation is accurate is not clear. Whether the results can be used to substantiate the findings cannot be known.

Additionally, Ethical Standard 9.02, Use of Assessments, states,

> (a) Psychologists administer, adapt, score, interpret, or use assessment techniques, interviews, tests, or instruments in a manner and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques.
>
> (b) Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation.
>
> (c) Psychologists use assessment methods that are appropriate to an individual's language preference and competence, unless the use of an alternative language is relevant to the assessment issues. (APA, 2010, p. 12)

Standard 9.06, Interpreting Assessment Results, states,

> When interpreting assessment results, including automated interpretations, psychologists take into account the purpose of the assessment as well as the various test factors, test-taking abilities, and other characteristics of the person being assessed, such as situational, personal, linguistic, and cultural differences, that might affect psychologists' judgments or reduce the accuracy of their interpretations. They indicate any significant limitations of their interpretations. (APA, 2010, p. 13)

Two general competence standards are applicable as well. Standard 2.01(b), Boundaries of Competence, states,

> Where scientific or professional knowledge in the discipline of psychology establishes that an understanding of factors associated with age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, or socioeconomic status is essential for effective implementation of their services or research, psychologists have or obtain the training, experience, consultation, or supervision necessary to ensure the competence of their services, or they make appropriate referrals. (APA, 2010, p. 5)

Finally, Ethical Standard 2.04, Bases for Scientific and Professional Judgments, indicates that psychologists should use only the best scientific and professional methods in their work. Unless linguistic equivalence is achieved to the highest standard possible, then the psychologist is in danger of failing to measure up to this standard.

Emic

The emic approach to test use has historically been at odds with the etic approach. An emic approach is consistent with an indigenous approach in that it is culture specific. In essence, tests are developed for particular groups under investigation without the need to expand them to other groups. It is limited in that a narrow understanding of a particular group does not increase one's broader understanding of psychological processes common to all individuals. However, we believe that more culture-specific tests are needed to gain a more robust understanding of diverse groups. Further theory development integrating both mainstream and indigenous psychologies will occur through increased development and recognition of culturally specific tests (Morris, Leung, Ames, & Lickel, 1999).

Although development and assessment of culture-specific tests has increased, a combined etic–emic approach to testing and assessment has recently received increased attention. Constructs derived indigenously are combined with local interpretations of universal constructs to offer a comprehensive measurement instrument relevant to a particular cultural context. Using an international example, the Chinese Personality Assessment Inventory (Cheung et al., 1996) is an instrument that combines both etic and emic perspectives. Local expressions of Chinese culture from a variety of China's regions served as the foundation for both culturally relevant and universal constructs. It overlaps with the Big Five scales, but a relational factor also emerged that is consistent with collectivistic cultures. It has great promise for future test development owing to the methodological approach taken, and it has been used in multiple regions of the world (Leong et al., 2010).

Additional Ethical Test Practices and Diversity

The APA Ethics Code has ethical practice standards that have relevance to diverse communities. These standards should be considered from a contextual framework. Some were mentioned earlier when discussing equivalence issues and two others are highlighted next. Although not explicitly stated, Standard 9.07, Assessment by Unqualified Persons, applies to those lacking sufficient cultural competence. For example, even culturally competent psychologists should be cognizant that not everyone with whom they work has the same level of cultural expertise. Colleagues should not be asked to administer, score, and interpret tests without proper understanding of their cultural context. When considering culture, this standard is also related to Standard 9.02, Use of Assessments. Standard 9.10, Explaining Assessment Results, becomes particularly salient when considering individuals whose second or third language is English and those who are unfamiliar with the purpose of testing. This standard is also related to Standard 9.03, Informed Consent in Assessments.

Although they are discussed in terms of school psychology assessments, Jacob and Hartshorne (2007) perhaps best summarized the ethical issues that arise from conducting broader culturally valid assessments. They determined that assessments should be multifaceted, comprehensive, fair, valid, and useful. As psychologists' understanding of cultural tests and assessments increases and becomes integrated into test development and use, they will feel comfortable using tests that cover these five issues, leading to greater ethical and cultural competence.

References

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Upper Saddle River, NJ: Pearson Education.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing* (Rev. ed.). Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, amended June 1, 2010)*. Retrieved from www.apa.org/ethics/code/index.aspx

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments. (1986).

Guidelines for computer-based tests and interpretations. Washington, DC: American Psychological Association.

Brislin, R. (1970). Back translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185–216. doi:10.1177/135910457000100301

Burlew, A. K. (2003). Research with ethnic minorities: Conceptual, methodological, and analytical issues. In G. Bernal, J. E. Trimble, A. K. Burlew, & F. T. L. Leong (Eds.), *Handbook of racial and ethnic minority psychology* (pp. 179–197). Thousand Oaks, CA: Sage. doi:10.4135/9781412976008.n9

Cheung, F. M., Leung, K., Fan, R., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese Personality Assessment Inventory (CPAI). *Journal of Cross-Cultural Psychology, 27*, 181–199. doi:10.1177/0022022196272003

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist, 66*, 593–603.

Dana, R. H. (2005). *Multicultural assessment: Principles, applications, and examples*. Mahwah, NJ: Erlbaum.

Distinguishing statements of ethical principles and regulatory guidelines. (2011). Retrieved from http://med.brown.edu/fogarty/codes.htm#disting

Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g.

Fisher, C. B. (2009). *Decoding the ethics code: A practical guide for psychologists* (2nd ed.). Thousand Oaks, CA: Sage.

Galton, F. (2003). Inquiries into human faculty and its development. In M. P. Munger (Ed.), *The history of psychology: Fundamental questions* (pp. 232–247). New York, NY: Oxford University Press. (Original work published 1883)

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191.

International Test Commission. (2005). *International guidelines on computer-based and Internet-delivered testing.* Retrieved from http://www.intestcom.org/itc_projects.htm

Jacob, S., & Hartshorne, T. S. (2006). *Ethics and law for school psychologists* (5th ed.). Hoboken, NJ: Wiley.

Jacoby, R., & Glauberman, N. (1995). *The bell curve debate: History, documents, opinions.* New York, NY: Random House.

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations.* Retrieved from http://www.apa.org/science/programs/testing/rights.aspx

Knapp, S., & VandeCreek, L. (2003). An overview of the major changes in the 2002 APA ethics code. *Professional Psychology: Research and Practice, 34*, 301–308. doi:10.1037/0735-7028.34.3.301

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., . . . Benton, S. A. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology, 60*, 725–739. doi:10.1002/jclp.20010

Leong, F. T. L., Leung, K., & Cheung, F. M. (2010). Integrating cross-cultural psychology research methods into ethnic minority psychology. *Cultural Diversity and Ethnic Minority Psychology, 16*, 590–597. doi:10.1037/a0020127

Leong, F. T. L., Qin, D., & Huang, J. L. (2008). Research methods related to understanding multicultural concepts. In J. K. Asamen, M. L. Ellis, & G. L. Berry (Eds.). *Handbook of child development, multiculturalism, and media* (pp. 63–80). Thousand Oaks, CA: Sage.

Meara, N. M., Schmidt, L. D., & Day, J. D. (1996). Principles and virtues: A foundation for ethical decisions, policies, and character. *Counseling Psychologist, 24*, 4–77. doi:10.1177/0011000096241002

Morris, M. W., Leung, K., Ames, D., & Lickel, B. (1999). Views from inside and outside: Integrating emic and etic insights about culture and justice judgment. *Academy of Management Review, 24*, 781–796.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 192–227). Baltimore, MD: Johns Hopkins University Press.

Schmitt, N., Golubovich, J., & Leong, F. T. L. (2010). Impact of measurement invariance on construct correlations, mean differences and relations with external correlates: An illustrative example using Big Five and RIASEC measures. *Assessment*. Advance online publication. doi:10.1177/1073191110373223

Schoenfeldt, L. F. (1989). Guidelines for computer-based psychological tests and interpretations. *Computers in Human Behavior, 5*, 13–21.

Schultz, D. P., & Schultz, S. E. (2011). *A history of modern psychology* (10th ed.). Belmont, CA: Wadsworth.

Spies, R. S., Carlson, J. F., & Geisinger, K. F. (2010). *Eighteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Takushi, R., & Uomoto, J. M. (2001). The clinical interview from a multicultural perspective. In

L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment* (2nd ed., pp. 47–66). San Francisco, CA: Jossey-Bass.

Turner, S. M., DeMers, S. T., Fox, H. R., & Reed, G. M. (2001). APA's guidelines for test user qualifications: An executive summary. *American Psychologist, 56*, 1099–1113. doi:10.1037/0003-066X.56.12.1099

Usunier, J. C. (1998). *International and cross-cultural management research*. London, England: Sage.

CHAPTER 16

THE IMPORTANCE OF EDITORIAL REVIEWS IN ENSURING ITEM QUALITY

Cathy Wendler and Jeremy Burrus

Appropriate editorial reviews are critical to develop and maintain an adequate supply of test items. Developing an adequate supply of items requires more than simply writing the items needed to build a specific number of tests. Test items may be lost during the review process as a result of several problems, including inappropriate editorial style, use of language, and level of language. As such, an editorial review plays an important role in creating high-quality test items.

Item reviews are integral to the development and maintenance of a testing program. As indicated in the *Standards for Educational and Psychological Testing* (American Educational Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), such reviews should be carried out by "expert judges" (see Standard 3, Test Development and Revision), but a definition of what constitutes an "expert" is not provided. The *Standards* also do not specify the criteria to be used when doing item editing or conducting a review. Through the years, several sources have provided guidelines to the practitioner regarding item writing (e.g., Haladyna & Downing, 1989; Haladyna, Downing, & Rodriguez, 2002; Krypsin & Feldhusen, 1974; Roid & Haladyna, 1982; Roid & Wendler, 1983). However, fewer sources have offered insight into how to conduct meaningful, efficient, and adequate item reviews (e.g., Baranowski, 2006; Schmeiser & Welch, 2006). Many of the guidelines for the item-reviewing process also apply to the item-writing process. However, although it is true that good item writing is essential to the development of high-quality tests, the process is incomplete without a separate and thorough item review.

ITEM REVIEWS AND VALIDITY EVIDENCE

Why are item reviews important? Overall item quality is established as part of the item review process by examining editorial quality, fairness across all subpopulations of examinees, and the appropriateness of item content. In addition, item reviews are an important step in providing validity evidence for the interpretations based on scores from the test.

A thorough item review is essential in determining whether test scores generalize to the target domain (Kane, 2006). That is, would performance on the test really provide information about how the respondent would perform on the construct in general? This assumption will not hold, however, if students who are not successful on the test would be successful in the real world. For example, a mathematical problem-solving item that involves a lot of reading may be difficult for an English language learner to answer correctly. In this case, the student may not be successful on the item not because he or she has difficulty with mathematical problem solving but because he or she has difficulty with reading English. Thus, the student's answer to the item bears little relation to how he or she would perform on a mathematics problem in the real world. In other words, the item is not valid for that individual. A properly conducted item review should locate potential threats to validity such as this one before

DOI: 10.1037/14047-016
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

they become a problem. Such a review provides content-related evidence of validity by ensuring that the items appropriately reflect the test design.

According to Kane (2006), providing evidence of content-related validity is closely related to the test development process. Claims as to the appropriate use and interpretation of test scores guide the development of the test and the items by defining the target domain and specifying the attribute of interest. Most frequently, the target domain is defined using content specifications or a test blueprint. The central focus of content specifications is to ensure that the test appropriately represents the content domain (Messick, 1993).

In testing programs in which multiple test versions are constructed and used, it is important that each test version reflects the test blueprint. This way, each form of the test appropriately measures what it should and thus allows judgments regarding the usability and meaningfulness of test scores across versions and provides fundamental content-related evidence of validity.

TYPES OF ITEM REVIEWS

To ensure quality items, reviews should routinely be conducted on each item. Most major test publishers routinely include multiple reviews as part of the item development process. These reviews can be placed into two categories: statistical and nonstatistical.

Statistical reviews follow item analysis and may include examining item difficulty levels, point-biserial correlations, differential item functioning indicators and, in the case of multiple-choice items, the distracters themselves. Items may be eliminated or revised on the basis of this review. A number of sources are available that discuss the use of item analysis as part of item development (e.g., Livingston, 2006; Chapter 7, this volume).

Nonstatistical reviews include content, editing, and fairness reviews (see also Chapter 17, this volume). Such reviews are critical to ensuring the quality (and, hence, the validity) of an item. However, despite the best reviews, there are times when an item's statistical properties may call the item's quality into question. For example, a distracter analysis, usually run as part of the item analysis, may indicate that one of the incorrect options in a multiple-choice item has been chosen by a high percentage of examinees. This especially becomes a problem when high-performing examinees choose an incorrect option along with lower performing examinees. Closer examination of the option may reveal that it overlaps or is slightly aligned with the correct option. In this case, the item may be classified as flawed even though the content, editing, and fairness reviews were appropriately accomplished.

Every review, be it statistical or nonstatistical in nature, is intended to evaluate particular aspects of a test question and, thus, is accomplished by experts with different knowledge and skills. However, the ultimate goal of all reviews is to ensure that the highest quality items are administered to examinees. The specific goals and the individuals responsible for conducting each review are summarized in Figure 16.1.

The focus of this chapter is on the editing review. However, because editing also occurs during the content review and may occur as part of a fairness review, these two reviews are also touched on.

Content and fairness reviews are generally carried out by assessment specialists or test developers. In some cases, members of expert panels may be involved in one or more of the stages of item review. The editing review is frequently conducted by an editor. Each review has a specific purpose and outcome, and performing a review efficiently and appropriately often requires that the reviewer receive specialized training.

Content Review

The first type of review is generally a content review. The goal of this review is to ensure that the item has a correct answer, that the content contained in the item reflects the test specifications, and that the item language and content is at the appropriate level for the population who will be taking the test. Obviously, having reviewers who are experts in the content being tested is important. In addition, because the person who wrote the item may have difficulty seeing flaws, especially if they are minor, the content review should be done by another expert, not by the item writer. Even the most experienced item writer may occasionally produce a flawed item, but more

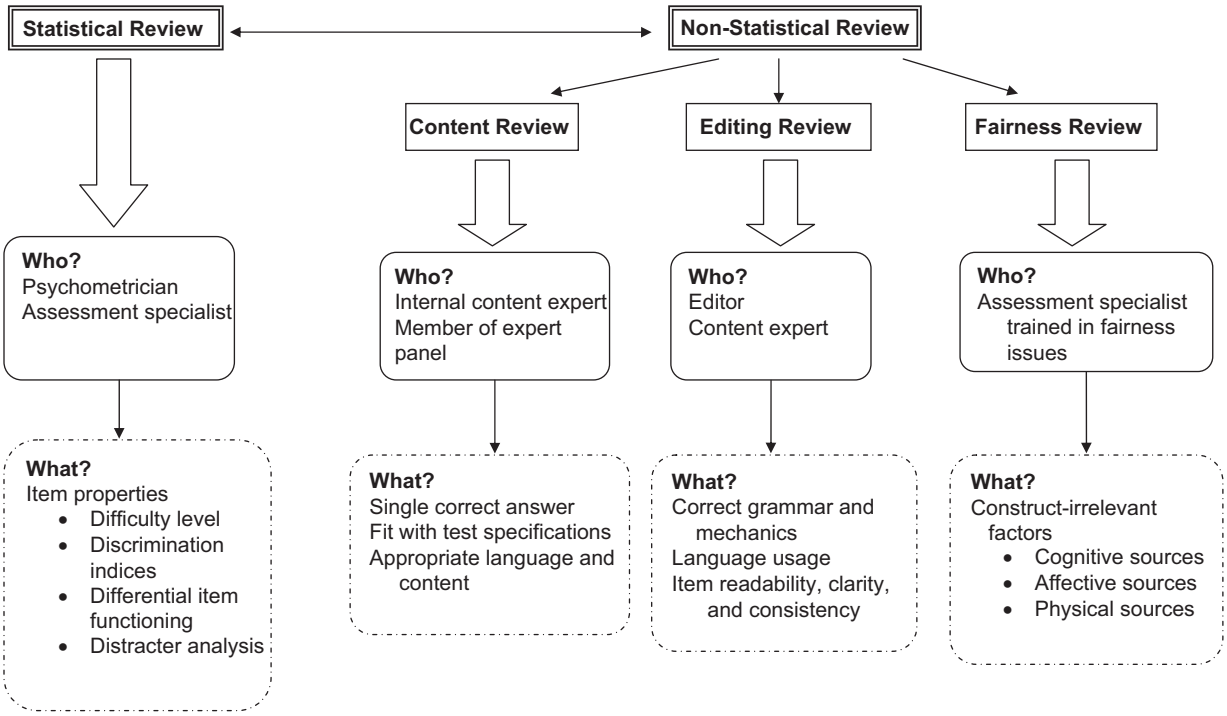

FIGURE 16.1. Goals and the roles of experts in item review.

important, most items can be improved by a good reviewer. Therefore, the reviewer's task is to not only capture flaws that would invalidate the item but to suggest improvements to the item.

The content review of standard multiple-choice items appears to be fairly straightforward, but it requires the reviewer to notice subtle things about an item. The reviewer should approach the item as would an examinee by reading the item and then choosing the correct response. If the reviewer, as expert, has difficulty locating the single best answer or is drawn to a particular distracter, it is clear that the item is flawed in some way. After the initial attempt, the reviewer should use a systematic approach to inspect the item as a whole and each of its parts. Exhibit 16.1 presents examples of questions that may routinely be considered as part of the content review.

Options in a multiple-choice item may be of two types: the key (correct response) or a distracter (incorrect response). When reviewing the options, additional considerations need to be kept in mind, depending on whether the option is the key or a distracter.

The content reviewer should agree that the option identified as the key by the item writer is the single best answer and determine whether there could be a situation in which the key might not be the best answer. In addition, the reviewer should examine each distracter to determine whether a case could be made by an examinee that a particular distracter is a plausible key. The wording and content provided in each distracter need to be reviewed to ensure that its content is appropriate enough to attract examinees who are not adequately prepared to answer the item. However, the distracter should not be incorrect because of a minor or unimportant detail that a knowledgeable examinee might miss.

Sometimes the content specialist knows little about writing items, and the item writer knows little about the content of the item. If this is the case, it is necessary to perform two checks: one for item content and one for technical item-writing issues. For example, it is important to ensure that the key does not stand out from the other distracters in any obvious way. A key that overlaps with the wording of the stem, that is more detailed, or that is noticeably different from the other options might go unnoticed by someone not familiar with the item-writing process. A check of the distracters by an individual experienced in item writing is needed so that distracters

Exhibit 16.1
Examples of Questions to Consider as Part of the Content Review

Item as a whole
- Does the item test the knowledge, skills, or abilities required by the test specifications in an appropriate manner for the intended testing population?
- Does the item appear to be within the range of difficulty for the intended testing population?
- Is there a better way to test the knowledge, skills, or abilities?
- Is the item confusing, unnecessarily difficult, or tricky?
- Could the item be shortened without losing clarity? Could the item be made easier to read?
- Are there references in the item that may cause it to become outdated?

Item stem
- Does the item clearly define the task or problem that the examinee must perform?
- Is unnecessary or misleading information contained in the stem?
- Can the stem be more clearly or concisely worded?

Item options
- Are the options reasonably parallel in structure?
- Do the options fit logically with the stem? Do they fit grammatically with the stem?
- Can any of the options be more clearly or concisely worded?
- Are the options in some reasonable order?
- Are quantitative options in order by size?
- Is any one option so inclusive that it eliminates another option from being considered?

Item key
- Does the reviewer's choice of the correct response agree with that of the item writer?
- Does the key answer the question as posed in the stem? Are there situations where the key may not be the best answer?
- Is there a better key among the options? Is there a better key than that provided in the options?
- Is the key obvious in an inappropriate way? (For example, is it very different from the other options? Is it the only option that repeats words from the stem? Is it more detailed than the other options?)

Item distracter
- Could a case be made that a particular distracter is a correct response?
- Is the distracter plausible enough to attract examinees that are misinformed or not properly prepared?
- Are there better distracters that are not listed?
- Is any distracter incorrect for a minor and unimportant reason that a knowledgeable examinee might miss?
- Does any distracter call attention to the key? For example, does the distracter state the opposite of the key?
- Do the distracters' content overlap with each other?

Note. Questions from Educational Testing Service, Princeton, NJ.

that, for example, state the opposite of the key are included in the range of the key or those that have content that overlap with each other are identified and revised.

Although item-writing guidelines generally call for the options to be placed in some type of reasonable order, there is no optimal placement for the item key. Some experts have suggested that a random placement of item keys is ideal (Kehoe, 1995), whereas others have believed that key placement has no impact on item validity (Bresnock, Graves, & White, 1989).

If a set of items is based on a reading passage, graph, chart, or other stimulus materials, the stimulus must also be examined as part of the content review. The common stimulus material should be inspected to ensure that it provides the necessary information to answer all of the items in the set. Other considerations for stimulus material in the form of a reading passage, short description, and so forth include whether it contains information that might be deleted or transferred to the stem of one of the items, whether it could be reorganized to be more logical or clear, and whether it could be worded more concisely or clearly. Stimulus materials in the form of a graph, chart, and so forth should be examined to ensure that they are properly labeled

and can be reproduced clearly in the test book or on a computer screen.

Finally, a review of the entire item set is necessary to ensure that the items work well together. Is answering the item without reading the stimulus unacceptable, or can it be allowed given the purpose of the item? Are items independent of each other, or do they provide clues that help to answer other items in the set? The method for delivering these items, either paper based or computer based, does not eliminate the need to review how the items and reading passage, graph, chart, or other stimulus material fit together within a set. Although some computer-adaptive tests may deliver only specific items from the set to an individual examinee, it is still important to establish the relationship of each item in the set to the reading passage, graph, chart, or other stimulus material.

Constructed response items, such as essay prompts, a short-answer task, and so forth, require a different review than multiple-choice items. As the task is written, the item writer should also construct the task's scoring criteria and any specific directions that accompany the task. The review of a constructed response item should include an evaluation of the task together with the directions and scoring criteria. This step will allow the reviewer to determine whether examinees are provided enough information to respond to the task in the intended way.

As with multiple-choice items, the directions for constructed response items also need to be complete, clear, and appropriate for the level of examinee being tested. Reviewing the scoring rubrics will also help determine whether they, as a whole, provide enough detail to score a response at the appropriate level. Scoring rubrics need to be worded clearly and phrased in ways that make them efficient to use.

The review of both multiple-choice and constructed-response items combines the content review and some elements of the editing review (see the next section). The reviewer should consider whether the entire task is appropriate for the purpose of the assessment, the population of examinees, and its relationship to the test specifications. The reviewer should also determine whether the phrasing of the task is complete, concise, and appropriate for the population to be tested.

Editing Review

During item editing, the item is reviewed for proper grammar and mechanics as well as other language issues such as usage, readability, clarity, and consistency. Although these issues are frequently caught during content review, it is during the editing stage that attention is focused only on grammar and language. It is important that item editing occur as part of the test development process and not as a final check once the test is assembled (Haladyna et al., 2002). Errors in grammar and spelling, unclear item stems and options, and inconsistency in language usage are examples of construct-irrelevant factors that interfere with examinees' ability to respond in a way that allows valid inferences to be made about their performance.

Item editing examines the item for grammatical and mechanical errors. It may also examine the level of language used and the consistency of language across a set of items. Ways to improve the clarity of the item should also be considered during this review. For example, items containing unnecessary negative phrases may be reworked so that only positive phrases are included. Items containing unique formats may be inspected to ensure that the special format is warranted. Exhibit 16.2 presents examples of questions that may routinely be considered as part of the editing review.

In addition, the item is examined to ensure that it meets particular style and format requirements. Style and format requirements are generally specific to each test and reflect the needs of the test content as well as the needs of the intended test population. For example, reading passages and items related to the passages that will be given to young examinees may contain slightly larger fonts, pictures, and simplified language compared with reading sets given to older examinees.

One result of poor item editing may be a nicely worded item that is wrong or unclear. That is, the editing of the item may have inadvertently changed the content of the stem or options. Because specialists involved in this review step may not be content experts, they must be cognizant of the impact of any change to the item. For example, item editing may rearrange the order of the options or reword the stem. If this occurs, it is important that a content

Exhibit 16.2
Examples of Questions to Consider as Part of the Editing Review

Item as a whole
- Is the item grammatically correct?
- Is the level of language used in the item appropriate for the intended testing population?
- Is the item confusing, unnecessarily difficult, or appear to be tricky?
- Could the item be shortened without losing clarity?
- Could the item be made easier to read?
- Does the item meet the appropriate style and format requirements?

Item stem
- Can the stem be more clearly or concisely worded?
- Is the level of language used in the stem appropriate for the intended testing population?
- Does the stem meet the appropriate style and format requirements?

Item options
- Are the options reasonably parallel in structure?
- Do the options fit grammatically with the stem?
- Can any of the options be more clearly or concisely worded?
- Are the options in some reasonable order?
- Do the options meet the appropriate style and format requirements?

Note. Questions from Educational Testing Service, Princeton, NJ.

expert determine whether the changes affect the quality of the item. If so, further revision of the item is required.

Fairness Review

The intent of the fairness review is to identify construct-irrelevant factors that may affect how members of a population subgroup respond to an item (Zieky, 2006). Construct-irrelevant factors may interfere with the ability to determine examinees' level of knowledge, skills, or abilities and draw into question the validity of inferences made from performance on the test. The three general sources of construct-irrelevant factors are (a) cognitive sources, such as difficult language, culturally specific language, and particular content or topics; (b) affective sources, such as inappropriate terminology or lack of diversity representation in items; and (c) physical sources, such as the use of charts, maps, and graphs that are irrelevant to the item construct; difficult-to-read fonts; and the use of specific letters or numbers that cause difficulty with alternate-format (e.g., Braille) tests. Removal of these factors is critical as part of good test development practice.

Fairness reviews are generally conducted using a prescribed set of guidelines. Zieky (2006) provided a full description of the rationale and procedures for the fairness review. In general, however, the review should determine whether the item contains language or content that could be offensive to or inappropriate for any population subgroup. The review should also determine whether any aspect of the item could be viewed as sexist, racist, or elitist. Finally, the review should determine whether any option would be considered unfairly attractive to members of any population group.

Resolving Disagreements

During the course of these reviews, changes are likely to be made to the item. Ensuring that any change made during the review process has not changed the intention of the item, made the key incorrect, or made a distracter a valid response is imperative. To accomplish this, a final review by a senior assessment specialist is needed. The senior assessment specialist must be an expert in the content being tested, recognize the impact of changes—even good changes—to the stem or options, and be able to communicate and negotiate if necessary with the item writer and all reviewers to reach a successful conclusion.

EXTERNAL REVIEWERS

Item reviews may be handled by assessment specialists within the company or institution building the assessment. At times, however, external reviewers may be required. For example, internal reviewers of items from occupational tests may have limited knowledge of the content of the test but understand the technical aspects of item writing. In these cases, outside experts are needed for the content review.

Some state educational testing contracts require that external panels of teachers review items before they are administered. These reviews may be in addition to those done by internal reviewers and often reflect the decisions made at the content and editing review stages.

Sometimes, specific knowledge of a particular area is required to do an adequate review, and the company creating the items may not have a sufficient number of staff in that area. In this case, test development panels of experts in the particular field may be brought together to perform the item reviews.

A methodology that is growing in use is that of the cognitive lab. The cognitive lab uses a "think-aloud" protocol that gathers input from examinees as they interact with test items. During the think-aloud protocol, respondents literally think aloud as they are reading and answering items (Willis, 1999). That is, they verbalize exactly what is going through their minds when they read an item and when they answer it. Interviewers typically do very little during this process besides read items to respondents and record their thoughts. Questions about items that may be answered during think-aloud protocols include the following: What does the respondent believe the question is asking? What do specific words and phrases in the question mean to the respondent? What types of information does the respondent need to recall to answer the question? What types of strategies are used to retrieve information? (Willis, 1999).

Interviews typically take place for no longer than an hour so as to avoid placing too much demand on the respondent (Willis, 1999). Large samples of subjects are not necessary to conduct a think-aloud procedure (typically, five to 10 are sufficient). After the think-aloud protocol has taken place, all examinees' comments are collected and categorized, and common themes are developed in a qualitative fashion. If issues are identified that may threaten the quality of an item, the item is edited and, optimally, another cognitive lab is conducted on the revised item. The cognitive lab approach is most useful as new, innovative items are developed or to ensure that examinees from subpopulations interpret the item in the intended manner.

UNIVERSAL DESIGN AND SUBPOPULATION CONSIDERATIONS

During item review, there is often a fine line between removing information that is considered to be construct irrelevant and ensuring that the construct is still being appropriately measured. One item development approach that attempts to mitigate this issue is that of *universal design* (UD), which refers to a particular style and test development process in which tests are designed and developed from the beginning to be accessible to the widest range of examinees, including individuals with disabilities and English language learners (ELLs; Johnstone, Altman, & Thurlow, 2006; Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008).

The four basic steps to developing a test under universal design are (a) conceptualization, (b) construction and review, (c) tryout, and (d) analysis. The step relevant to this chapter is construction and review.

During the item construction and review step, items are first written, and then edited, typically by a group of experts in the field of interest. During item writing and reviewing, particular attention should be paid to the guidelines for writing universal design items: item clarity, language usage (such as avoiding the use of negative stems), alternate-format considerations, and avoiding content that depends on knowledge or experience that certain groups of individuals may lack (e.g., that not all examinees can see or hear or when examinees have not had the same cultural experiences). Stimulus materials containing complex features may be difficult to script for a reader or for an audio format of the test. Passages or other content that assume specific sensory abilities may disadvantage examinees with particular impairments. For example, examinees with poor vision, color deficiency, or a cognitive impairment may have difficulty recognizing certain features that are included in a figure, such as labels that are small, printed on shaded areas, or vertical or slanted. Certain letters and numbers should not be used as well. For example, in mathematics problems, using the letters *A* through *J* as variables should be avoided because of their confusion with Braille numbers.

Furthermore, the stem of an item should be concise and contain clear signals about how examinees

should direct their attention (e.g., "in the second sentence").[1] Complex descriptions should be avoided in the stem and the options. Not only may this increase the reading load unnecessarily, but audio formats of a test usually present the stem followed by Choice A, then present the stem followed by Choice B, and so forth. As a result, the listening load becomes very burdensome for the examinee.

Universal design can be especially useful in designing assessments that may be taken by ELLs. ELLs are examinees whose proficiency in English is less than that of a native speaker of English. As such, several factors may affect their test performance (Pitoniak et al., 2009), including (a) language factors (different linguistic backgrounds, varying levels of English proficiency, and varying levels of proficiency in the native language), (b) educational background factors (varying degrees of formal schooling in the native language and in English), and (c) cultural factors (acculturation to the U.S. mainstream).

Item reviewers should not assume that ELLs have had previous exposure to or experience with responding to any particular item type. Constructed-response items should explicitly indicate what type of response is acceptable (paragraph, complete sentence, list, mathematical equation, etc.) and the criteria that will be used to evaluate the response. Directions should be clear and concise to minimize the potential for confusion.

Reviewers need to find cases in which an item can be clarified or simplified to make the language more accessible. However, it may not be possible or acceptable to simplify language that is part of the construct being measured. Steps leading to accessible language include (a) using vocabulary that is widely accessible to examinees and avoids colloquial and idiomatic expressions, unduly difficult words, and words with more than one meaning; (b) maintaining as simple a sentence structure as possible; (c) avoiding the use of negatives and constructions that use *not* in the stem or options; and (d) using a context familiar to a wide range of students (e.g., a school-based context is often more accessible to ELLs than a home-based context).

In addition, reviewers should pay attention to formatting issues. Font type, font size, line breaks in paragraphs, and test directions should be reviewed. ELLs who have learned to read in another language may have differing familiarity with text that reads in a particular orientation (e.g., right to left vs. left to right). Consistency in placing elements such as pictures, graphs, page numbers, and so forth can greatly improve readability for ELLs.

LESSONS FROM PSYCHOLOGICAL RESEARCH

Research in psychology has identified several issues that are relevant to writing and reviewing items (Schwarz, 1999; Schwarz & Oyserman, 2001). This research has posited that the way in which tests and items are constructed, including the types of options presented, influences the answers that an examinee provides and thus can introduce construct-irrelevant variance.

Although these issues apply mostly to psychological assessments (including those used in schools, usually involving self-reports of behavioral frequency and attitudes), several are also relevant to educational assessments. A complete review of these issues is beyond the scope of this chapter. Schwarz (1999) and Schwarz and Oyserman (2001) provided more detailed information on these issues.

ENSURING QUALITY ITEMS

Appropriate and thorough reviews of items for content, language, and fairness, along with knowledge of the impact of the scale and other item features, are essential to creating quality items. Item reviews help identify and remove construct-irrelevant factors to ensure that the best measurement of the construct is being achieved. Item reviews are the first step in ensuring that valid inferences are made from a test.

However, the key to producing valid, quality items is not through reviews but by having strong item-writing procedures in place from the beginning. Item reviews are integral to ensuring high-quality items and are critical steps in good item

[1]Note that some state programs do not allow any reference to graphics in the item stem.

development. Reviewers can catch and correct errors in content and language and identify features that draw into question the fairness of the item, but reviewers can also inadvertently change content or language in inappropriate ways and miss errors that classify an item as flawed. Strong item development guidelines, well-trained and expert assessment specialists, and a predefined process all help ensure that items are written with high quality. In such cases, item reviews become a matter of validating the item or making minor tweaks to it. It is most important to stress writing items correctly and not rely on reviews to fix them.

References

American Educational Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Baranowski, R. A. (2006). Item editing and editorial review. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 349–357). Hillsdale, NJ: Erlbaum.

Bresnock, A. E., Graves, P. E., & White, N. (1989). Multiple-choice testing: Questions and response position. *Journal of Economic Education, 20*, 239–245. doi:10.2307/1182299

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 51–78. doi:10.1207/s15324818ame0201_4

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–333. doi:10.1207/S15324818AME1503_5

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multi-method approaches to item review. *Educational Measurement: Issues and Practice, 27*, 25–36. doi:10.1111/j.1745-3992.2008.00112.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kehoe, J. (1995). *Writing multiple choice test items*. Retrieved from http://www.ericfacility.net/ericdigests/ed398236.html

Krypsin, W. J., & Feldhusen, J. F. (1974). *Developing classroom tests: A guide for writing and evaluating test items*. Minneapolis, MN: Burgess.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421–441). Hillsdale, NJ: Erlbaum.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Phoenix, AZ: Oryz Press.

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English-language learners*. Princeton, NJ: Educational Testing Service.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing*. New York, NY: Academic Press.

Roid, G. H., & Wendler, C. L. (1983, April). *Item bias detection and item writing technology*. Paper presented at the 67th annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: Praeger.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105. doi:10.1037/0003-066X.54.2.93

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication and questionnaire construction. *American Journal of Evaluation, 22*, 127–160.

Willis, G. B. (1999). *Cognitive interviewing: A "how to" guide*. Research Triangle Park, NC: Research Triangle Institute. Retrieved from http://appliedresearch.cancer.gov/areas/cognitive/interview.pdf

Zieky, M. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Hillsdale, NJ: Erlbaum.

CHAPTER 17

FAIRNESS REVIEW IN ASSESSMENT

Michael J. Zieky

Fairness review is an inspection of test items (test questions) and stimuli to identify materials that the reviewer believes may result in less valid measurement for some demographic groups of test takers. Differences across groups in cognitive, affective, and physical variables that are unrelated to the purpose for testing may contribute to invalid score differences among groups of test takers. Test materials must be written and reviewed with care to avoid such problems. Therefore, the general principles of fairness review are based on the need to avoid sources of invalid score differences that may affect different groups of test takers in different ways. The general principles are instantiated by specific guidelines that may vary from country to country.

PURPOSE

The purpose of this chapter is to explain how to help ensure the fairness of tests by using fairness reviews to identify potentially invalid aspects of items and stimuli that may impede the appropriate measurement of test takers in different demographic groups. This chapter is written for people who write, review, or edit test items; people who assemble or review tests; people who commission test development services; and people who are interested in fairness in assessment. No prior knowledge of measurement or statistics is required.

OVERVIEW

The chapter places fairness review in the context of the test development process and discusses definitions of *fairness, fairness review,* and a few related terms. The link between fairness and validity is clarified. The compelling rationale underlying the beginnings of fairness review is explained, and the resulting rapid growth of fairness review is briefly noted. Guidelines for fairness review based on the general principle of avoiding invalid score differences among groups are then described in detail, including additional guidelines commonly used in testing children. The chapter closes with a discussion of the effects of fairness review and suggestions for procedures to guide the application of fairness review.

FAIRNESS REVIEW IN CONTEXT

Although the focus of this chapter is limited to fairness review, it is important to note that fairness review alone is not sufficient to ensure that tests are fair. Test developers must pay attention to fairness throughout the test development process and not simply tack on a fairness review at the end. According to the *ETS Standards for Quality and Fairness* (Educational Testing Service [ETS], 2002, p. 12), test developers must "address fairness in the design, development, administration, and use" of tests. For example, diversity of input in test planning is required. The people who set test specifications and the people who write items must follow guidelines for fairness to avoid the inclusion of unfair material in the first place. Fairness reviews should be a fine tuning to identify and remove subtle problems that have inadvertently been introduced. Fairness review

DOI: 10.1037/14047-017
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

should not be the first time fairness concerns have been addressed in the test development process. During test administrations, accommodations should be provided, as appropriate, for test takers with special needs. (For information on testing people with special needs, see Volume 3, Chapter 18, this handbook.) After a test has been administered and when sample sizes are sufficient, empirical analyses of item and test characteristics should be carried out for various groups of test takers. (See Chapter 7, this volume, for information on analyses.) Finally, the consequences of test use for different groups of test takers should be investigated. (For a discussion of validity, including the consequences of testing, see Chapter 4, this volume.)

DEFINITIONS OF FAIRNESS

Before discussing fairness review, it is necessary to define what *fairness* means in the context of assessment because it is impossible to carry out effective reviews for an undefined attribute. Fairness is not a unitary concept. The authors of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) refused to be limited to a single definition of the term because "*fairness* is used in many different ways and has no single technical meaning" (p. 74). (See Chapter 13, this volume, for more information about the *Standards for Educational and Psychological Testing*. See Camilli, 2006, and Volume 3, Chapter 27, this handbook, for a discussion of the various meanings of fairness in assessment.)

Most psychometric definitions of *test fairness* are not helpful in reviewing tests for fairness before their use, not only because the definitions are based on the results of using completed tests, but also because the definitions are often contradictory. (See, e.g., the various definitions of fairness in Cleary, 1968; Cole, 1973; Darlington, 1971; and Linn, 1973.)

The popular, but incorrect, definition of *fairness* as the equality of average scores across groups of test takers is similarly of no use for a fairness review. To illustrate that score differences alone are not proof of bias, consider tape measures. Even though the average heights of men and women differ, tape measures are unbiased. (For more on this issue, see, e.g., Cole & Zieky, 2001, and Thorndike, 1971.) Even though group score differences are not proof of bias, they are often a basis for legal challenges to the use of tests. (See Chapter 38, this volume, and Volume 3, Chapter 25, this handbook, for discussions of legal issues in testing.)

FAIRNESS AND VALIDITY

Score differences between demographic groups do not necessarily mean that a test is unfair. However, group score differences caused by factors unrelated to what the test is intended to measure do mean that a test is unfair. The authors of the *Standards for Educational and Psychological Testing* clearly stated that "*bias* in tests and testing refers to construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees" (AERA et al., 1999, p. 76; for more information on detecting bias, see Chapter 8, this volume).

The definition of *bias* in the *Standards* ties fairness directly to validity. Messick (1989) described validity in terms of the "adequacy and appropriateness of interpretations and actions based on test scores" (p. 13). A *construct* is the set of knowledge, skills, or other attributes a test is intended to measure. A *construct-irrelevant component,* therefore, is a contaminant that degrades measurement of the intended construct, causes invalid variance (differences among scores), and interferes with the appropriate interpretation of test scores. A test is unfair if invalid aspects of the test cause significant score differences among defined groups of test takers.

Although fairness is closely related to validity, fairness can be differentiated from validity. Fairness is decreased if an invalid source of variance affects test takers in different groups in different ways or in different amounts. If a source of construct-irrelevant variance, such as an ambiguous item, affects all groups of test takers equally, the test is less valid than it might otherwise be, but it is still fair. If, however, a source of construct-irrelevant variance, such as unnecessarily difficult language in a mathematics

test, affects some group of test takers, such as English language learners, more than it affects English-proficient test takers, then the test is less valid for the English language learners. Less valid measurement for some groups of test takers makes the test less fair. However, if the English language learners received lower scores because they were less able in mathematics, then the score difference would be valid and, therefore, fair.

DEFINITION OF FAIRNESS REVIEW

Fairness review is an inspection of test items and the stimuli on which items are based to identify construct-irrelevant components that may result in invalid score variance for identifiable groups of test takers. Because fairness review is carried out before a test is administered and is often applied to individual items and stimuli before a test is even assembled, the effects of bias on the scores of identifiable groups are not known during fairness review. Therefore, fairness review is based on the identification of plausible (as defined by the fairness review guidelines in use) sources of construct-irrelevant variance. Fairness review should be used before tests are administered to minimize the exposure of test takers to potentially unfair materials, even if empirical indicators of fairness such as differential item functioning are used after the test has been administered. (See Dorans, 1989, and Zieky, 1993, for explanations of differential item functioning and its uses. See Chapter 7, this volume, for more information on empirical indices of item quality, and Volume 3, Chapter 27, this handbook, for psychometric perspectives on test fairness.)

BEGINNINGS OF FAIRNESS REVIEW

The current focus on the identification of construct-irrelevant components of items makes fairness review a means of enhancing validity rather than a means of enforcing political correctness. Fairness review began, however, with the simple belief that an item is unfair if it appears unfair. If a reviewer believed that an item would alienate, offend, anger, or otherwise disconcert test takers in some defined group, the item was considered unfair. That belief is unsophisticated and makes the judgment of fairness highly subjective, yet the belief remains widespread.

The rationale for a fairness review is clear, whether the focus is on construct-irrelevant components of items or on items that merely appear unfair. Test developers want to create tests that are fair for all identified groups in the population of test takers. Therefore, if reviewers can identify unfair items through inspection, then test developers must do fairness reviews and remove or revise any unfair items. Moreover, compared with many other aspects of test development, fairness review is inexpensive and easy to arrange and can be accomplished quickly.

Beginning in the late 1960s and continuing to the present, test developers, publishers, school districts, and government agencies have made fairness reviews more and more common and more and more comprehensive. (See Ramsey, 1993, and Zieky, 2006, for information on the history of fairness review.) After about a third of a century of growth, fairness review had become so prevalent and so inclusive that Ravitch (2003) complained, "What began with admirable intentions has evolved into a surprisingly broad and increasingly bizarre policy of censorship that has gone far beyond its original scope" (p. 4).

To the extent that fairness review is limited to the identification of construct-irrelevant components of test materials, it is a means of improving test quality rather than a form of censorship. To the extent that fairness review limits construct-relevant test material on the basis of the reviewer's idiosyncratic predilections, it may become a form of censorship. Whether fairness review is a "bizarre policy of censorship" or a good-faith effort to ensure that tests are as fair as possible for all who take them is clearly open to debate and may vary from test to test. What is not arguable, however, is that fairness review has become standard operating procedure for test developers and that it has an important influence on the contents of tests put together by major publishers.

SUPREMACY OF VALIDITY

The focus of fairness reviewers should be on the construct-irrelevant components of tests. Anything that is important for valid measurement is

necessarily construct relevant and, therefore, acceptable. Tests of specific subject matter such as anatomy, U.S. history, biology, psychology, and so forth may appropriately include material that would be unacceptable in a test of general skills, such as reading comprehension, for which no specific content is required. For example, tests for licensing physicians may include items on abortion procedures that would be considered offensive or excessively controversial in other contexts.

GROUPS

Theoretically, fairness review applies to any demographic group of test takers. In practice, however, considering all groups is impossible. Therefore, attention is given primarily to groups that have historically been discriminated against because of differences from the dominant groups in age, ethnicity, gender, national or regional origin, native language, physical or mental abilities, race, religion, sexual orientation, or socioeconomic status. The particular groups of primary concern for fairness review vary from place to place. For example, in the United States, Hispanic people would be considered a group requiring special attention, but in Chile, they would not.

FAIRNESS REVIEW GUIDELINES

Fairness reviews should be based on written guidelines to help ensure that all relevant aspects of fairness are considered; to provide guidance to test designers, item writers, and item editors; and to reduce subjective differences among reviewers.

Consistency of Guidelines

Because all fairness reviewers have the same goal of freeing tests from content and images that are considered unfair, it is no surprise that the reviewers in similar cultures tend to develop and use similar sets of guidelines. Ravitch (2003) pointed out that the bias guidelines promulgated by educational publishers, test development companies, states, and scholarly and professional organizations overlap so extensively "that it is difficult to disentangle them" (pp. 32–33). After reviewing a large number of different sets of guidelines, she remarked, "All test and textbook bias guidelines start to look alike" (p. 53). The guidelines do tend to be worded similarly because they focus on the same concerns, and there are, for example, relatively few ways to indicate that Americans of African descent should be described as *African American* or *Black* rather than as *Negro*, or that *man* should not be used to refer to all human beings.

Given their great similarities in content and wording, there is little need to discuss guidelines from multiple sources. The *ETS Guidelines for Fairness Review of Assessments* (ETS, 2009a) serves as the source for guidelines discussed in the remainder of this chapter. For more than 30 years, various editions of the *ETS Guidelines* have been used by hundreds of test developers to structure reviews of a wide variety of tests administered throughout the world to test takers from the elementary grades through graduate school. The guidelines state that "ETS allows use of the guidelines by all who wish to enhance the fairness of their tests" (ETS, 2009a, p. 1).

Sources of Construct-Irrelevant Variance

Fairness review is a search for construct-irrelevant sources of score variance that may differentially affect identifiable groups of test takers. Mean differences across groups in cognitive, affective, and physical variables may serve as significant sources of construct-irrelevant variance, if items are not written appropriately.

Cognitive sources of construct-irrelevant variance stem from differences, unrelated to what the test is intended to measure, in the knowledge bases of different groups. Affective sources stem from differences among groups in the materials that elicit emotions that may interfere with the ability to respond appropriately to test material. Physical sources stem from differences in the visual, aural, or motor abilities of some groups in perceiving and responding to test material.

General Principles of Fairness Review

Those sources of construct-irrelevant variance lead directly to the three general principles of fairness review (ETS, 2009a, p. 4):

1. Avoid cognitive sources of construct-irrelevant variance. . . .

2. Avoid affective sources of construct-irrelevant variance. . . .
3. Avoid physical sources of construct-irrelevant variance.

These general principles of fairness review are anchored in validity theory and apply to all tests and all test-taking populations. The particular instantiations of the principles, however, may vary across cultural contexts (e.g., from country to country or even from region to region within a country). For example, an image of women playing soccer in short-sleeved shirts and short pants may be seen as a positive, healthy image in some cultures yet be considered offensive in other, more conservative cultures. The guidelines discussed later are appropriate for tests used primarily in the United States and in countries with a similar culture and values. (See ETS, 2009b, for advice on generating locally appropriate fairness review guidelines for use in different countries. See Volume 3, Chapters 9 and 26, this handbook, for information on assessing individuals from different cultures.)

Avoiding Cognitive Sources of Construct-Irrelevant Variance

For a wide variety of sociocultural reasons (e.g., average differences in interests, experiences, family backgrounds, environments, schooling, cultural expectations), different groups can develop average differences in knowledge of various topics. For example, men in the United States are likely to know more about military topics than women do, and African American test takers are likely to know more about the struggle for civil rights in the United States than White test takers do. (See Volume 2, Chapter 27, this handbook, for examples of gender-related differences.)

If groups differ in construct-relevant knowledge, then the differences in scores that ensue are valid and fair. If, however, knowledge of some topic other than what the test is intended to measure is required to answer an item, then the ensuing differences in scores are less valid and less fair. For example, an item designed to measure division with decimals may ask about the number of nickels in $5.85. People unfamiliar with U.S. coins may be able to do the division but may not know that a nickel is .05 of a dollar. The item is unfair for such people because construct-irrelevant knowledge that they lack is required to answer the item. (Note that the item would be fair if the construct included knowledge of U.S. coins.)

Linguistic difficulty. A common cognitive source of construct-irrelevant variance is unnecessarily difficult language in items. Language that is excessively difficult may affect all test takers, but people who are not native speakers and people with language-related disabilities are most at risk. Rules for accessible language are beyond the scope of this chapter, but many useful references are available. (See, e.g., the *Publication Manual of the American Psychological Association* [APA, 2010] and the *Chicago Manual of Style* [University of Chicago Press, 2010]; for more information on language issues in assessment, see Chapter 21, this volume, and Volume 3, Chapter 10, this handbook.)

Test developers should avoid construct-irrelevant sources of linguistic difficulty. It is important to make the language of an item no more difficult than is required for valid measurement. Test developers should, whenever possible, use straightforward syntax and common words. They should avoid construct-irrelevant specialized terminology used in subject areas such as agriculture, finance, law, science, and technology. They should also avoid construct-irrelevant regionalisms, words that are much more common in some regions of the country than in others.

Construct-relevant sources of linguistic difficulty are, of course, acceptable. For example, construct-relevant knowledge of many subjects includes knowledge of the specialized vocabulary used in the subject.

Problematic topics. Some topics, such as religion, sports, and the dominant culture of the country, are likely to be cognitive sources of construct-irrelevant variance because groups differ in their familiarity with those topics.

Test developers should not require specialized knowledge of sports, such as the number of people on a hockey team, to answer items about other topics. Even though women are more involved in sports than they were previously, the topic is still a

common cause of construct-irrelevant variance between male and female test takers.

The topic of religion is often a problem because of the false assumption that all test takers are familiar with the most common religion in the country in which the test was developed. Test developers should not require knowledge of religion to answer an item unless the purpose of the item is to measure such knowledge. If religious knowledge is intertwined with the tested subject (as Christianity is with medieval European history), it is important to test only those aspects of religion learned in the study of the subject rather than those learned in the study of the religion.

Test developers should not require construct-irrelevant information that is likely to be known only by people familiar with the government, history, holidays, institutions, laws, locations, public figures, and so forth of the country in which the test was developed. For example, assuming that all test takers would know how Halloween is celebrated in the United States would be wrong.

Avoiding Affective Sources of Construct-Irrelevant Variance

Strong emotional reactions to test materials may impede test takers' performance or may lead test takers to believe that their performance has been impeded. If construct-irrelevant components of test materials are more likely to cause strong emotional reactions for some groups than for other groups, test fairness may be reduced.

Some topics have proved so likely to cause negative reactions that it is best to avoid them entirely unless they are important for valid measurement. For example, in the United States, abortion, abuse of people or animals, contraception, euthanasia, harmful experimentation on people or animals, rape, Satanism, torture, and witchcraft are topics best avoided unless important for valid measurement. No list of problematic topics, however, can ever be complete because current events can always add new topics. Consider, for example, the emotional impact of a highly publicized terrorist attack. Any test materials that are related to salient aspects of the attack, such as the location, may be sources of affective construct-irrelevant variance.

Other topics need not be avoided but must be treated carefully to minimize their potential emotional impact.

Advocacy. Test developers should avoid test materials that advocate for one position on a controversial issue because such materials can disadvantage test takers who hold opposing views. It is important to avoid construct-irrelevant materials about conflict between ethnic or religious groups, including people closely associated with one of the sides in such conflict. Some types of items require test takers to evaluate an argument or defend a point of view, so some controversial material may be necessary. In such cases, test developers should use the least controversial material that will allow valid measurement.

Evolution. Evolution is a clear example of a topic that engenders strong emotions in some groups, but evolution is also a key concept in biology. Test developers should allow the topic of evolution when it is important for valid measurement, as in biology tests, and avoid the topic when it is not important for valid measurement, as in tests of logical reasoning. By avoiding the topic of evolution only when it is not relevant to the tested construct, test developers are not taking a stand against evolution. They are simply avoiding an unnecessary source of negative emotional reactions for some test takers.

Group differences. Discussion of group differences, particularly innate differences, is likely to be an affective source of construct-irrelevant variance. When such discussions are construct relevant, test developers should limit them to findings about objectively measured traits supported by research. They should avoid generalizations about more subjective issues such as group differences in courage, industriousness, physical attractiveness, or quality of culture.

Pain and death. When dealing with topics such as accidents, death and dying, illnesses, natural disasters, self-destructive behavior, suffering, and violence, test developers should avoid gruesome, shocking, graphic details unless they are important for valid measurement, as in the case of tests used to license emergency medical technicians.

Religion. Test developers should avoid any focus on religion in general, or on any specific religion, unless the focus is important for valid measurement. If mention of a religion can be construed as positive, adherents of other religions may believe the material is advocating the mentioned religion. If it can be construed as negative, adherents of the religion may feel attacked. Passing references to religion are acceptable if they are neutral and factual.

Sex. Many test takers consider explicit descriptions of human sexual activity to be offensive. Test developers should avoid such descriptions unless they are construct relevant, as in some tests for medical personnel. If tests are to be used with people from very conservative cultures, it is best to avoid construct-irrelevant images or descriptions of people in revealing clothes or of people engaged in immodest behavior.

Stereotypes. Stereotypes are likely to anger the people being stereotyped. Test developers should not reinforce stereotypes in test materials. Showing people in traditional activities (such as a woman caring for a child) is acceptable in an item as long as the traditional activities are balanced by non-traditional activities in other items. Showing only traditional activities for group members reinforces stereotypes. Language in which stereotypes are embedded, such as *man-sized job,* should be avoided. (For information about stereotypes, see Volume 2, Chapter 25, this handbook.)

Terminology for groups. Derogatory names for groups should never be used. Test developers should try to use the labels that group members prefer. When groups are first mentioned, it is preferable to use group names as adjectives rather than as nouns. For example, *Black students* is preferable to *Blacks.* Later references can use the group names as nouns sparingly.

Some older group names (e.g., *Negro, colored, Oriental*) should be limited to quotations from literary and historical material or in the names of organizations. *African American* and *Black* are the preferred terms for Americans of African descent. *People of color* is acceptable but *colored people* is not. Test developers should use *Asian American* or, preferably, more specific terms such as *Chinese American* or *Korean American* for Americans of Asian descent. Test developers should use *Hispanic American* or *Latino American* (*Latina American* for women) to refer to Americans of Spanish descent or, preferably, more specific terms such as *Mexican American* or *Guatemalan American. Bisexual, gay, lesbian,* and *transgender* are appropriate terms. Test developers should not use *queer* except in reference to academic courses with that title. *White* and *Caucasian* are both acceptable, but *White* is preferred.

An important rule is to use parallel terminology for women and men. For example, if men are indicated by title and last name (e.g., *Mr. Smith*), women should be indicated by title and last name as well. If women are indicated by first name only, men should be indicated by first name only as well. If women are described by role in the family (e.g., *wife, mother*), men should be similarly described (e.g., *husband, father*).

The terms *boy* and *girl* should be used only for people younger than age 18. Test developers should not use *he* or *man* to refer to all people. In general, compound words that include the word *man* such as *fireman, foreman,* and *mankind* should be avoided. Gender-neutral terms such as *firefighter, supervisor,* and *people* are preferable. Terms for roles such as *doctor, nurse,* and *scientist* include both men and women. It is not appropriate to add gender identifiers such as *male nurse* or *female scientist.*

For people with disabilities, the focus should be on the person rather than on the disability. For example, *a person who is blind* is a better descriptor than *a blind person* when the person is first mentioned. Later references can refer to a *blind person.* Test developers should avoid references to groups using the noun form of the disability as in *the deaf* or *the blind* except in the names of organizations or in quotations from literary or historical material.

It is best to be objective and neutral in descriptions of people with disabilities. Both overly "correct" terms such as *special* or *challenged* and overly negative words such as *afflicted* or *victim* are best avoided. Test developers should refer to people as *normal* or *abnormal* only in biological or medical contexts.

Representation of diversity. If a test includes people, test takers should, ideally, be able to see people

like themselves in the test. For example, women may feel excluded if all of the people mentioned in a test are male. Although it is not possible to do so in any single test form, test developers should try to represent the different groups in the test-taking population across test forms, to the extent allowed by the subject matter. For example, a reading comprehension test would offer more opportunities to include various groups than would a chemistry test. To the extent allowed by the subject matter, members of all groups included in the tests should be shown in a variety of social roles. For example, test developers should not make all the managers male and all the assistants female.

Avoiding Physical Sources of Construct-Irrelevant Variance

It is important to avoid unnecessary physical barriers to test performance. Some physical barriers, however, are necessary for valid measurement. For example, if a test is supposed to measure an aspiring language teacher's ability to detect students' errors in pronunciation, the ability to hear students' speech is necessary for valid measurement, even though it is a physical barrier to people who are deaf or hard of hearing. Some physical barriers, however, are not necessary, such as the use of a very small font. This barrier could be removed with no loss in the ability of the item to measure the intended construct.

Following are examples of the types of physical barriers in items and stimuli that should be removed or revised unless they are construct relevant (ETS, 2009a, pp. 37–38):

- visual stimuli (charts, maps, graphs, pictures) that are primarily decorative rather than informative, or are more complex or difficult to read than necessary;
- visual stimuli in the middle of a block of text;
- use of fine distinctions in shade or color to mark important differences, or text that contrasts poorly with the background;
- excessively small or decorative fonts, or lines of text that are slanted or curved;
- the use of letters that look alike (e.g., O and Q) or sound alike (e.g., S and X) as labels in the same diagram;
- unclear or low-volume recordings; and
- excessive scrolling to access material in computer-based tests.

Additional Guidelines for Testing Children

Many jurisdictions that commission test development services for schools use additional guidelines to help ensure the appropriateness of test content for children. In general, the additional guidelines deal with material that might upset or frighten children, might offend or exclude groups of children, or might model bad behavior. The additional concerns for children tend to be extensions of the second general principle for fairness review, "Avoid affective sources of construct-irrelevant variance" (ETS, 2009a, p. 4).

Even though the guidelines for testing children are often particularly strict, any material required for valid measurement may be used. For example, social studies tests may appropriately include the topic of slavery, even though that topic would be excluded from a test of reading comprehension.

Upsetting materials. Test developers should avoid topics likely to upset children, such as serious illnesses, divorce, domestic violence, parental loss of jobs, disputes among family members, fights, social ostracism, bullying, and arguments between students and teachers. Pests (e.g., rats, lice) and frightening animals (e.g., venomous snakes, scorpions) are generally inappropriate topics in tests for children.

Offensive materials. Test developers should avoid potentially offensive topics such as swearing, gambling, smoking, drinking alcohol, and using illegal drugs. They should also avoid construct-irrelevant references to any deity. Celebrations that may exclude groups of children who do not participate in them, such as birthday parties and religious holidays (including Halloween and Valentine's Day), are best avoided in tests.

It is important to be strictly neutral about controversial issues such as gun control, global warming, energy policy, environmentalism, unions, welfare, political candidates, and so forth in test materials. A

focus on discrimination, sexism, or racism is to be avoided unless such a focus is construct relevant.

Inappropriate behavior. Test developers should avoid showing models of bad behavior among children such as lying, cheating, stealing, cutting school, or doing dangerous things. Test materials should not appear to be cynical about values that most people want their children to have, such as being honest, working hard, being patriotic, and the like.

EFFECTS OF FAIRNESS REVIEW

It has long been known that the relationship between reviewers' judgments of item fairness and empirical findings of item bias is, at best, weak. (See, e.g., Bond, 1993; Cole, 1981; Plake, 1980; Tittle, 1982.) In an early version of its fairness review guidelines, ETS (1987) stated that the importance given to fairness review did not rest on "a measurable relationship between material considered offensive by some test takers and the scores of test takers" (p. 4). Ramsey (1993), in a discussion of fairness review (then called *sensitivity review*), wrote, "There is no promise to anyone that the sensitivity review process will raise the scores of, say, minority test takers" (p. 384).

What fairness review does promise is a good-faith effort to try to ensure that any mean score differences across demographic groups are based on construct-relevant sources of variance. In addition to being fair, tests must be perceived to be fair. No amount of data will convince test takers and score users that a test is fair if they find content that is unnecessarily offensive or insulting, reinforces stereotypes, uses derogatory labels for their group, and so forth. Fairness review removes construct-irrelevant material perceived to be unfair without demanding empirical proof that the material will actually cause unfair group differences.

PROCEDURES FOR APPLICATION OF FAIRNESS REVIEW

In the real world, tests are made within constrained schedules and with limited budgets. Managers, subject-matter specialists, and fairness reviewers may have competing priorities and different views of fairness. To reduce conflict, it is crucial to select or develop written guidelines for fairness of test content and to distribute them to all involved in the test development process. Test developers should be trained to apply the guidelines consistently. Only trained people should do fairness reviews. Idiosyncratic reviews should be discouraged by requiring reviewers to explain why any challenged material is out of compliance with one or more of the written guidelines.

No set of written guidelines can cover all possible situations. Judgment will always be required in the application of fairness review guidelines. What is considered fair depends not only on the importance of the material for the measurement of the intended construct, but also on the characteristics of the test takers. People will disagree about the fairness of test materials. Therefore, it is necessary to establish a system, such as a steering committee, to resolve disputes between item writers and fairness reviewers who cannot reach agreement on their own. Test developers should maintain written records of reviews and the resulting revisions, keep track of issues that lead to disputes, and use the information to help clarify the guidelines when they are revised. Because views of fairness change over time, it is useful to review the guidelines at least once every 5 years to determine whether they need to be revised.

CONCLUSION

Fairness is an aspect of validity. To the extent that fairness review succeeds in making tests fairer by removing construct-irrelevant content that affects different groups in different ways, it makes tests more valid. There is, therefore, a firm psychometric foundation for fairness review that extends well beyond the enforcement of political correctness. In addition to making tests more valid and more fair, fairness review decreases the perception that tests are unfair. The potential benefits of fairness review far exceed the relatively small amount of time and expense that it adds to the test development process.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–280). Hillsdale, NJ: Erlbaum.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (pp. 221–256). Washington, DC: Praeger.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10*, 237–255. doi:10.1111/j.1745-3984.1973.tb00802.x

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36*, 1067–1077. doi:10.1037/0003-066X.36.10.1067

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369–382. doi:10.1111/j.1745-3984.2001.tb01132.x

Darlington, R. B. (1971). Another look at "culture fairness." *Journal of Educational Measurement, 8*, 71–82. doi:10.1111/j.1745-3984.1971.tb00908.x

Dorans, N. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel–Haenszel method. *Applied Measurement in Education, 2*, 217–233. doi:10.1207/s15324818ame0203_3

Educational Testing Service. (1987). *ETS sensitivity review process: An overview*. Princeton, NJ: Author.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author. Retrieved from http://www.ets.org/about/fairness

Educational Testing Service. (2009a). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Author. Retrieved from http://www.ets.org/about/fairness

Educational Testing Service. (2009b). *ETS international principles for fairness review of assessments: A manual for developing locally appropriate fairness review guidelines in various countries*. Princeton, NJ: Author. Retrieved from http://www.ets.org/about/fairness

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43*, 139–161.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40*, 397–404. doi:10.1177/001316448004000217

Ramsey, P. (1993). Sensitivity review: The ETS experience as a case study. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.

Ravitch, D. (2003). *The language police: How pressure groups restrict what students learn*. New York, NY: Knopf.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8*, 63–70. doi:10.1111/j.1745-3984.1971.tb00907.x

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 31–63). Baltimore, MD: Johns Hopkins University Press.

University of Chicago Press. (2010). *Chicago manual of style* (16th ed.). Chicago, IL: Author.

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zieky, M. J. (2006). Fairness review. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Mahwah, NJ: Erlbaum.

PART II

TYPES OF TESTING

CHAPTER 18

OBJECTIVE TESTING OF EDUCATIONAL ACHIEVEMENT

Michael C. Rodriguez and Thomas M. Haladyna

Educational achievement tests can be classified in many ways. Perhaps the most global classification is the distinction between *standardized* and *teacher-made* tests. This distinction clarifies the degree to which content, format, administration, and scoring are standardized or left to the discretion of the teacher or testing agent (see Volume 3, Chapter 16, this handbook). Another distinction in educational tests is the format of the test items. More commonly used tests are classified as *objectively* or *subjectively* scored. The extent to which tests are objective or subjective is usually a function of the type of scoring required given the item or task type. These terms are components of a continuum rather than a dichotomous choice. Although some items might be considered strictly objectively scored because a correct response is known and responses may be machine scored (or scored with a key), other items may be subjectively scored because they require human scoring or a complex scoring guide such as a rubric, with which many correct (or partially correct) responses are possible. Even so, many forms of subjectively scored items, such as short-answer constructed-response (CR) items or even essays, may be scored with simple to complex computer scoring engines. Perhaps most performance assessments, with complex multidimensional aspects, are most unlikely to be objectively scored (see Chapter 20, this volume).

When focused more specifically on the nature of the test items or tasks themselves, additional classification schemes are possible. The two categories used in this chapter are selected-response (SR) and CR items. For SR items, options are available from which to select responses, whereas for CR items, the respondent must construct or produce a response because none are available for selection. Similarly, items may be considered fixed-response items (SR) versus free-response or production items (CR). In practice, most educational test developers use the terms *multiple choice* (*selected response*) and *constructed response* because these formats are predominant in current educational tests. However, *multiple choice* as a label refers to a specific class of selection items, and so it is not all encompassing. Within the two classifications of SR items and CR items, there are many possible formats. In the class of CR items are formats that could be objectively scored, including the cloze, fill-in-the-blank, grid-in response, and short-answer formats and even some experiments or tasks that result in specific products. More subjectively scored production items include demonstrations, debates, essays, exhibitions, interviews, observations, and more complex performance tasks, portfolios, projects, and research papers (see Haladyna & Rodriguez, in press, for a comprehensive review of these types).

The focus of this chapter is on objective testing of educational achievement and thus items that lend themselves to objective scoring (see Chapter 19, this volume, for additional perspectives). This category includes a wide range of SR items and those CR items that can be objectively scored. In the realm of SR items, the typical formats include multiple choice (MC) and variants, true–false (TF), and matching formats. Recently, other innovative computer-enabled

DOI: 10.1037/14047-018
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

selection-type item formats have been introduced; the chapter describes these formats briefly. The chapter focuses on item writing and task development. For more general considerations of test development, see Chapter 9, this volume. The approaches here are most directly applicable to teacher-made tests but could be used in a more systematic approach to the design of standardized tests. The development of standardized tests requires additional systematic procedures, including formal editorial review (see Chapter 16, this volume); item analysis (Chapter 7, this volume); fairness review (Chapter 17, this volume); perhaps norming, scaling, and equating (Chapter 11, this volume); and in some cases, standard setting (Volume 3, Chapter 22, this handbook).

ITEM-WRITING GUIDELINES

Many of the item-writing guidelines currently recognized as effective or essential to quality test development have been implemented in practical settings, often in connection with large-scale testing programs. However, few of these guidelines have been studied empirically. The most notable advances occurred in the early 1900s when written exams such as the U.S. Army Alpha in 1917 and college entrance exams from the College Board during the 1920s began using MC item formats (see DuBois, 1970, for a complete, if somewhat dated, discussion of the history of testing). MC test items are able to elicit brief and targeted responses in a specific content domain. Because of their simplicity, in both construction and administration, MC items presented a practical alternative to other item formats such as constructed, written, or oral responses. In the first edition of *Educational Measurement*, Ebel (1951) discussed the research and development of item-writing techniques during the first half of the 20th century, highlighting the prevalence of the MC item. He also lamented the limited attention to the science of item writing.

Since the early 1900s, advances in technology, most notably the advent of computer-based testing, have lead to more innovative item formats, ones that include features such as video, sound, and interactivity with the examinee. In many ways, these advances have resulted in improved measurement because innovative formats allow test items to better represent the constructs they are written to measure. For example, in testing of health science students and medical professionals, skill in performing complicated surgical procedures can be tested using video or even three-dimensional simulations of specific anatomical regions (Shanedling, Van Heest, Rodriguez, Putnam, & Agel, 2010). In the testing of mathematics or statistics, test items may integrate spreadsheets or plots to better gauge students' understanding. In reading, students can interact with text to identify words, phrases, or passages that serve specific literary functions and even reorder phrases and passages to modify (e.g., clarify or correct) the sequencing of information. Such tasks potentially become more authentic and construct sensitive.

Researchers in other arenas have also offered sage advice about item development in the context of objective testing of educational achievement. These arenas include universal design and accessibility. Rodriguez (2009) provided a review of psychometric considerations regarding the development of alternate assessments and reviewed several articles of a special edition of *Peabody Journal of Education* devoted to accessibility. The goal of accessibility in item and test development is to support the measurement of a given construct under unique learner conditions, particularly those with cognitive, physical, or emotional impairments. These issues are largely beyond the work covered in this chapter, but additional information can be found in Volume 3, Chapters 17 and 18, this handbook.

This chapter is intended to cover the essential aspects of item development in objective testing to enhance the quality of tests and test scores. The first two sections contain an overview of the common formats of SR and CR items that lend themselves to objective testing. The third section reviews item-writing guidelines for both types of item formats. The final section includes an introduction to item validity and includes a model of gathering qualitative validity evidence from the item development process and a review of the promise of innovative item types to enhance the measurement of a construct.

SELECTED-RESPONSE FORMATS

We guess that many SR formats have not been studied but have been created through novel modification of a small number of typical formats. The most common formats for SR items are based on the ubiquitous MC item (as presented by Haladyna, Downing, & Rodriguez, 2002). The following examples are representative, but many new types of items are being developed by innovative item writers, especially taking advantage of technology and computer-enabled testing.

- *Conventional MC:*
  When it comes to describing the distribution, the standard deviation tells us
  A. where most of the scores are located.
  B. if the distribution is normal.
  C. how far the scores are spread out.
- *Alternate choice:*
  If a distribution of raw scores is positively skewed, converting to *T* scores will result in what type of distribution?
  A. Normal
  B. Positively skewed
- *True–false (dichotomous choice):*
  True or false: If the item difficulty index is .70, then 30% of examinees answered the question correctly.
- *Multiple true–false:*
  Consider the following actions that may affect test score validity evidence. Determine whether each is true or false.
  1. Adding more test items of similar quality improves test score validity.
  2. Increasing the sample size will increase criterion-related validity correlations.
  3. Obtaining a sample with more test score variability increases criterion-related validity correlations.
  4. Eliminating items with poor item–total correlations (discrimination) will improve content-related validity evidence.
- *Matching:*
  Match each term on the right with the description on the left.

1. score stability	A. systematic error
2. attention-deficit disorder	B. random error
3. content alignment	C. item difficulty
4. *p* value	D. item discrimination
5. item–total correlation	E. reliability evidence
	F. validity evidence

- *Context-dependent item set:*
  An anonymous standard item analysis report was found online that appears to be a cumulative report for an exam in a specific course. This exam has been completed by 327 students. The total number of items is 50. Refer to this item analysis report [not shown in this example] when answering the following items.
  1. Which item is the easiest? _____
  2. Which item should be revised into a TF item? _____
  3. Which item has the best discrimination? _____
  4. Identify one item that has the best example of effective distractors. _____
  5. Identify one item that is most likely to have two correct answers. _____
- *Complex MC (use of this type is not recommended; see later discussion):*
  Which are norm-referenced interpretations of test scores?
  1. John's score is 3 standard deviations above the class mean.
  2. Mary answered 80% of the items correctly.
  3. Eighty percent of the class scored above a *T* score of 45.
  4. The average math score for Arlington High is equal to the district average.
  5. Antonio is proficient in fifth-grade reading.
     A. 1 and 3.
     B. 2, 3, and 5.
     C. 2 and 5.
     D. 1, 3, and 4.
     E. All 5.

CONSTRUCTED-RESPONSE FORMATS

CR items have a much wider variety of formats, and they have not been consistently organized in the literature. CR items are different from SR items because they require the examinee to generate or construct a response. Osterlind and Merz (1994) and Haladyna (1997) described more than 20 formats for CR items. They also included some forms

of performance assessment. These tasks require much more extensive scoring rubrics, substantially more time, and more planning and preparation time, so they are typically not adaptable to on-demand testing or objective scoring.

CR item formats can be classified in several ways because they can differ in response modes allowed for various item formats or in the scoring processes for various item formats. In large-scale achievement tests, one typically finds grid-in items, short-answer items, and essay formats. In many cases, responses to such item formats can be objectively scored, particularly through the use of automated scoring (see, e.g., Attali & Burstein, 2006).

ITEM-WRITING GUIDELINES

Many educational measurement textbooks are available to students, researchers, and measurement practitioners, and nearly all contain one or more chapters on item writing. Some chapters are designed to be comprehensive reviews of item development, including Chapters 12, 13, and 14 in the *Handbook of Test Development* (Downing, 2006; Welch, 2006; and Sireci & Zenisky, 2006, respectively) and, more generally, Chapters 9 and 16 in *Educational Measurement* (Ferrara & DeMauro, 2006, and Schmeiser & Welch, 2006, respectively). Entire books have also been devoted to item writing, including *Writing Test Items to Evaluate Higher Order Thinking* (Haladyna, 1997), *Constructing Test Items* (Osterlind, 2002), *Developing and Validating Multiple-Choice Test Items* (Haladyna, 2004), and *Developing and Validating Test Items* (Haladyna & Rodriguez, in press). These resources are important for more in-depth preparation. To facilitate the presentation of item-writing guidelines, the guidelines most appropriate for MC items are presented first, many of which apply to other SR and CR formats. A small number of important guidelines are also available for TF and matching SR formats. Guidelines for CR formats are also presented.

Multiple-Choice Item-Writing Guidelines

The first research-based taxonomy of MC item-writing guidelines was developed by Haladyna and Downing (1989a), followed by a summary of the available empirical evidence (Haladyna & Downing, 1989b). This taxonomy was revised to include additional empirical evidence and a meta-analytic review of some of that evidence (Haladyna et al., 2002) and further refined to combine related components (Haladyna & Rodriguez, in press). Most of the guidelines are based on principles of good writing, logical reasoning, and lessons learned from practice; very few are based on empirical evidence. Item-writing guidance covers four elements of MC items: content, formatting and style, writing the stem, and writing the options.

Content concerns are possibly the most important. The subject-matter expert provides the leadership for writing a successful item. Items must be carefully written to include important relevant content and cognitive skills. These guidelines are largely based on the logic and experience of item writers and on examinee reactions. Aside from some general research on clarity and appropriate vocabulary use, no empirical studies of these item-writing guidelines exist. Guidelines about content concerns include the following:

1. Base each item on one type of content and cognitive demand.
2. Use new material to elicit higher level thinking.
3. Keep the content of items independent of one another.
4. Avoid overly specific and overly general content. Test important content.
5. Avoid opinions unless qualified.
6. Avoid trick items.

Formatting and style concerns are based on good writing practice. Some empirical evidence has supported general use of most item formats (Haladyna et al., 2002), whereas others, such as the complex MC format, appear to introduce greater difficulty that may be unrelated to the construct being measured in most settings. At least 17 studies have examined the difference in difficulty of items formatted as a complete question stem versus an open stem that is completed by the options. No difference was discernible in this body of research. The complex MC (Type K) item format is generally much more difficult (by .12 on average, in 13 studies), yielding lower reliability (by .15 on average, in 10

studies) in most settings. These guidelines for formatting and style include the following:

7. Format each item vertically instead of horizontally.
8. Edit and proof items.
9. Keep linguistic complexity for the group being tested appropriate.
10. Minimize the amount of reading in each item. Avoid window dressing.

Additional evidence is available regarding Guidelines 9 and 10 (see Volume 3, Chapter 17, this handbook).

Writing the stem is another area in which the empirical evidence is limited. With the exception of negatively worded stems, which empirical results suggest should rarely be used (Haladyna et al., 2002), these guidelines are extensions of style concerns, specifically applied to the item stem. In experimental research on using negatively worded stems, at least 18 studies resulted in an average slight increase in item difficulty, with a small subset suggesting a loss of reliability (average decrease in coefficient alpha of .17, nonsignificant because of a small sample, $n = 4$). Although the work of Abedi and others has provided evidence to support these guidelines, their work was not intentionally designed to test the validity of specific item-writing guidelines (Abedi, 2006, 2009; Hess, McDivitt, & Fincher, 2008). The guidelines regarding writing the stem include the following:

11. State the central idea in the stem very clearly and concisely. Avoid repetitious wording.
12. Word the stem positively, and avoid negatives such as *not* or *except*.

Writing the choices is the area in which the most research studies have been done. This area has 15 specific guidelines, but only three have been studied multiple times empirically and experimentally (13, 18, 20a). The one guideline that has received the most attention in the research literature regards the number of options in a MC item. Rodriguez (2005) conducted a comprehensive meta-analysis of 80 years of research on this topic and concluded that three options are sufficient, if not optimal. Also notable in the item-effects research literature, at least 17 studies have examined the effect of making the correct option longer than the distractors, a common error found in teacher-made tests (Haladyna et al., 2002). This characteristic tended to make items easier, increasing the item *p* value by .06 on average, with a dramatic drop in validity coefficients (−.26, nonsignificant because of the small sample of four studies). Finally, the use of *none of the above* has been found to make items slightly more difficult by an average of .04 (in 57 studies) and may have the effect of slightly decreasing item discrimination (by .03, *ns*). Guidelines for writing the choices include the following:

13. Use only options that are plausible and discriminating. Three options are usually sufficient.
14. Make sure that only one of the options is the right answer.
15. Vary the location of the right answer.
16. Place options in logical or numerical order.
17. Keep the content of options independent; options should not be overlapping.
18. Avoid using *none of the above, all of the above,* or *I don't know*.
19. Word the options positively; avoid negatives such as *not*.
20. Avoid giving clues to the right answer.
    a. Keep the length of options approximately equal.
    b. Avoid specific determiners including *always, never, completely,* and *absolutely*.
    c. Avoid clang associations—options identical to or resembling words in the stem.
    d. Avoid pairs or triplets of options that clue the test taker to the correct choice.
    e. Avoid blatantly absurd, ridiculous options.
    f. Keep options homogeneous in content and grammatical structure.
21. Make all distractors plausible. Use students' typical errors in writing distractors.
22. Avoid humorous options.

Other Selected-Response Item-Writing Guidelines

There exists, in the many educational measurement textbooks, a rather disorganized set of guidelines for other SR item formats. Here we present some of the more common recommendations for these

additional SR formats. All of the MC item-writing guidelines presented earlier apply, as appropriate, to the other SR formats.

Alternate-choice items are MC items with only two options. In such a case, the two options should contain a correct response and a distractor. In all MC items, the plausibility of the distractor is important, but in the alternate-choice item, it seems more critical. The distractor should be selected to announce the existence of an error in thinking, a misconception, or a misunderstanding. Such items can serve important diagnostic (formative) purposes.

TF items remain an important tool in classroom assessment (as well as in some personality assessments, such as the Minnesota Multiphasic Personality Inventory, and interest inventories). As with all SR items, it is important that TF items do not focus on trivial facts. Nitko (2001) provided many examples of well-written TF items with a checklist for judging the quality of such items. Frisbie and Becker (1991) provided a summary of 17 textbook author recommendations regarding the use of TF items. A couple of recommendations from that review include the following:

1. Balance the number of true and false statements.
2. Use simple declarative sentences.
3. Write items in pairs to help identify inherent ambiguity (while only using one in a given test).
4. When the statement is a comparative one, put the comparison directly into the statement.

Multiple TF (MTF) items are a hybrid of MC and TF item formats. Each option in the MC item becomes a TF statement, with respect to the stem. Each MTF item may have two or more such TF options, and the number may vary from item to item within a test. A good strategy for initially constructing MTF items is to start with a MC item. Poorly functioning MC items (those with multiple true responses) can be converted into effective MTF items. As with the other SR item formats, one must balance the number of true and false correct responses across MTF items and make sure that at least some of the options within each MTF item are true and false.

Matching items are also an important and common tool in classroom assessment. A matching item has explicit instructions on how to complete the matching, a list of premises, and a list of responses. The instructions are particularly important because the test developer must concretely explain the basis for matching. In one sense, these items function similarly to MC items, because each premise is like a MC stem and the list of responses serves as the options. It is important to be consistent with the MC guidelines. A few unique guidelines are needed for matching items.

1. The matching exercise should be homogeneous.
2. All responses should be a plausible response to each premise.
3. The list of premises and options should not be overwhelmingly long, given the examinees' ability.
4. The number of premises should be different than the number of responses. This can be accomplished by allowing responses to be used multiple times or including responses that do not match to any premise.

Context-dependent items constitute a testlet that functions as a small test within a larger test because a set of test questions relates to a common stimulus (reading passage, graphical display, etc.). Haladyna and Rodriguez (in press) provided guidance for the use of context-dependent item formats, with suggestions for connecting the context to the items, thus facilitating deeper item development considerations. For example, items should be written that depend on the context rather than on being independent—the answer is only known if the specific context provided in the test is understood. It is important to place the items in close proximity to the context material to facilitate this dependence, on the same or opposing page if possible. In general, the item-writing guidelines for the other formats apply to the items used in the context-dependent format.

Constructed-Response Item-Writing Guidelines

Guidelines for writing CR items are less developed and, when writing CR items, measurement specialists have less agreement on what is important. Also, fewer research studies have investigated the importance of CR item-writing guidelines. Osterlind and

Merz (1994) proposed a taxonomy for CR items, largely based on the work of cognitive psychologists. This taxonomy contained three dimensions: (a) the type of reasoning competency used, including factual recall, interpretive reasoning, analytical reasoning, and predictive reasoning; (b) the nature of cognitive continuum used, including convergent thinking and divergent thinking; and (c) the kind of response yielded, including open product and closed product formats, producing 16 combinations. The first two dimensions address cognitive processes. The third dimension addresses the kinds of responses possible. Closed-product formats are those that allow few possible response choices, possibly scored with a relatively simple scoring key; open-product formats permit many more choices, requiring scoring with more elaborate rubrics, potentially allowing unanticipated innovative responses. With respect to objective testing, the closed-format tasks are more likely to be successful. Here, the concern is the degree to which responses can be objectively scored.

Most major testing companies develop CR item-writing guides to direct the work of their item writers. For example, the Educational Testing Service produces several large-scale tests that include CR items (e.g., National Assessment of Education Progress and Advanced Placement exams). These testing programs have resulted in a large body of research on the quality of CR items, but not much of this is published. The Educational Testing Service's *Guidelines for Constructed-Response and Other Performance Assessments* (Baldwin, Fowles, & Livingston, 2005) provides good advice. These guidelines are general, forming a basis from which more specific guidelines can be developed for specific testing programs. Baldwin et al.'s (2005) guidelines describe what item development specifications must contain, including specification of the domain of knowledge and skill, issues related to cultural and regional diversity, response modes and conditions of testing, timing, the number and types of tasks, and scoring processes.

Hogan and Murphy (2007) summarized a functional set of item-writing guidelines based on their review of 25 textbook authors. We note that these guidelines reflect many of the principles provided in the MC item-writing guidelines presented earlier. However, they are not entirely consistent with the allowances made in the Educational Testing Service guidelines (Baldwin et al., 2005) and result from some disagreement among textbook authors (e.g., the Educational Testing Service and Hogan & Murphy [2007] disagreed about whether students should be allowed choice in responding to CR items). Moreover, the potential cost of human scoring requires additional consideration of the use of CR items. Most important, the CR item should require the kinds of thinking that are not easily obtained from other SR formats. In some cases, this requires the allowance of novel responses, making the items less likely to be objectively scored.

The guidelines for closed-product CR items can similarly be written to cover content, format, and style concerns, writing the directions and stimulus, and a general context concern. Without greater description and an explanation of their bases, Haladyna and Rodriguez (in press, Table 11.1) provided the following guidelines:

Content Concerns

1. Clarify the domain of knowledge and skills or tasks to be tested.
2. Determine the desired cognitive demand that the item is supposed to elicit for a target set of test takers.
3. Choose the format that has the highest fidelity for the intended content and cognitive demand.
4. Assure construct comparability across tasks.

Formatting and Style Concerns

5. Edit and proof the items.
6. Pilot items and test procedures.

Writing the Directions/Stimulus

7. Clearly define directions, expectations for response format, and task demands.
8. Provide information about scoring criteria.
9. Avoid construct-irrelevant task features.

Context Concerns

10. Consider cultural and regional diversity and accessibility.
11. Ensure that the linguistic complexity is suitable for intended population of test takers.

Technology-Enabled Innovative Items and Tasks

Computer-based testing has provided an opportunity for a variety of new item formats, and innovations continue. Some have pointed to the greater accessibility of these innovations for students who struggle with typical SR items, whereas others have argued that such innovations allow one to tap the target construct more directly—in both cases improving validity. Sireci and Zenisky (2006) summarized 13 computer-enabled item formats that they believed also enhance construct representation, potentially reducing construct-irrelevant variance. A strong validity argument depends on the degree to which the construct is sampled and represented in the items and tasks presented on a test. A couple of these innovative formats are described here.

The extended MC item is typically associated with a reading passage, in which each sentence in the passage plays the role of an option that can be selected as an appropriate response to specific questions. Such questions can ask about the main idea of a paragraph, for which the examinee highlights the appropriate sentence exemplifying the main idea directly in the reading passage. The options are the sentences within the reading passage itself rather than out-of-context statements or an idiosyncratic rephrasing of the main idea in the typical MC item.

Other formats allow examinees to connect ideas with various kinds of links (dragging and connecting concepts), to sort concepts, or to order information. The computer environment enables a wide range of innovative response processes, including correcting sentences with grammatical errors; correcting or editing mathematical statements; completing written statements or mathematical equations; and producing, completing, or altering graphical models, geometric shapes, or trends in data. Computers provide a wide range of possibilities that potentially mimic real-life activities (e.g., what an architect might do) in a way that is not easily facilitated by paper and pencil.

Postsecondary and professional exam programs have taken advantage of the computer environment to enable item and response innovations. The GRE and the Test of English as a Foreign Language programs have experimented with innovative response options and have studied their impact on item and test score quality. Bennett, Morley, Quardt, and Rock (1999) investigated the use of graphical modeling for measuring mathematical reasoning in the GRE General Test. As an example, examinees plotted points on a grid and used a tool to connect the points. Examinees agreed that the graphing items better represented their potential success in graduate school, but they preferred traditional MC items (which is a result commonly found when comparing MC and CR items in other settings). Bennett et al. found highly reliable item scores that were moderately related to the GRE General Test quantitative total score.

ITEM VALIDATION

In measurement and testing, *validity* is frequently defined as the extent to which evidence supports a specific interpretation or application of scores from a test (see Chapter 4, this volume). Current definitions of validity vary across fields; however, in educational testing, most have agreed with the framework described in the *Standards for Educational and Psychological Testing* (AERA et al., 1999): "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). The *Standards* describes validation as the process of gathering evidence to achieve these goals, including evidence related to the construct, test content, response processes, internal structure, relations to other variables, and intended and unintended consequences.

In all cases, validation is an ongoing process, and the most important sources of validity evidence are those that are most closely related to the immediate inferences and proposed claims psychologists make regarding test results. Validity evidence is important to gather at all stages of test design, administration, analysis, and reporting. When considering item development, multiple forms of validity evidence can be gathered to support the use of specific items. A model to provide qualitative validity evidence for item quality was proposed by Downing and

Haladyna (1997). These sources of evidence include the following:

- content definition, found in documentation of item content selection methods;
- test specifications, found in the systematic link of test content to test blueprint;
- item writer training, found in training materials, methods, written materials, and sample items;
- item-writing principles, found in compliance with adopted item-writing rules and procedures;
- verification of item content, found in evidence of the cognitive classification system used for items and content expert reviews of items, including expert credentials;
- item editing, found in the editing process, results, and editor credentials;
- item review, found in the bias or sensitivity review process, results, and reviewer credentials;
- item tryout, found in pretest or pilot test examinee characteristics and results of item functioning;
- key verification, found in the policy and procedures for key verification and documentation of results; and
- security plan, found in the test security plan and monitoring procedures.

SUMMARY

Although the quality of items is clearly important, the research on item writing is sparse. All of the important decisions that are made on the basis of test scores, including placement, advancement, admissions, certification, and licensure, make test score quality of high importance. This necessity is, of course, a concern about validity. To enhance the validity of interpretations and uses of test scores, it is necessary to ensure that the examination consists of high-quality items.

In his 2009 NCME presidential address, Reckase (2010) proclaimed that test items are complicated. He likened them to little poems:

> A constrained literary form that requires careful choice of words and clear communication in a limited space. . . . There seems to be a belief that anyone can be a good item writer. . . . It would be better to identify people who have demonstrated good item writing skills, rather than expect that with minimal training anyone can do this critical creative job. For many years, I have thought it would be nice to honor great item writers in the same way that we honor other great writers. (p. 4)

This chapter has addressed validity at its core, through reviews of common SR and CR item formats, common and evidence-based item-writing guidelines, the promise of innovative item types, and qualitative validity evidence that can be gathered in the item development processes. These areas all address aspects of validity, because they are intended to improve the measurement quality of items, the basic building blocks of measures.

References

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Erlbaum.

Abedi, J. (2009). English language learners with disabilities: Classification, assessment, and ccommodation issues. *Journal of Applied Testing Technology, 10*(2). Retrieved from http://www.testpublishers.org/assets/documents/Special%20issue%20article%202.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Attali, Y., & Burstein, J. (2006). Automated scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment, 4*(3), 1–30.

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments.* Princeton, NJ: Educational Testing Service.

Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (1999). *Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning* (ETS RR-99-21). Princeton, NJ: Educational Testing Service.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Mahwah, NJ: Erlbaum.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 61–82. doi:10.1207/s15324818ame1001_4

DuBois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn & Bacon.

Ebel, R. L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 185–249). Washington, DC: American Council on Education.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., p. 324). Westport, CT: Praeger.

Frisbie, D. A., & Becker, D. F. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education, 4*, 67–83. doi:10.1207/s15324818ame0401_6

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Boston, MA: Allyn & Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 37–50. doi:10.1207/s15324818ame0201_3

Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 51–78. doi:10.1207/s15324818ame0201_4

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–333. doi:10.1207/S15324818AME1503_5

Haladyna, T. M., & Rodriguez, M. C. (in press). *Developing and validating test items.* New York, NY: Routledge.

Hess, K., McDivitt, P., & Fincher, M. (2008). *Who are the 2% students and how do we design items and assessments that provide greater access for them? Results from a pilot study with Georgia students*. Paper presented at the 2008 CCSSO National Conference on Student Assessment, Orlando, FL. Retrieved from http://www.nciea.org/publications/CCSSO_KHPMMF08.pdf

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education, 20*, 427–441. doi:10.1080/08957340701580736

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Osterlind, S. J. (2002). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). New York, NY: Kluwer.

Osterlind, S. J., & Merz, W. R. (1994). Building a taxonomy for constructed-response test items. *Educational Assessment, 2*, 133–147. doi:10.1207/s15326977ea0202_2

Reckase, M. D. (2010). NCME 2009 presidential address: "What I think I know." *Educational Measurement: Issues and Practice, 29*, 3–7. doi:10.1111/j.1745-3992.2010.00178.x

Rodriguez, M. C. (1997, April). *The art and science of item writing: A meta-analysis of multiple-choice item format effects.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*, 3–13. doi:10.1111/j.1745-3992.2005.00006.x

Rodriguez, M. C. (2009). Psychometric considerations for alternate assessments based on modified academic achievement standards. *Peabody Journal of Education, 84*, 595–602. doi:10.1080/01619560903241143

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., p. 324). Westport, CT: Praeger.

Shanedling, J., Van Heest, A., Rodriguez, M. C., Putnam, M., & Agel, J. (2010). Validation of an online assessment of orthopedic surgery residents' cognitive skills and preparedness for carpal tunnel release surgery. *Journal of Graduate Medical Education, 2*, 435–441.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Erlbaum.

Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–327). Mahwah, NJ: Erlbaum.

CHAPTER 19

OBJECTIVE PERSONALITY TESTING

Samuel E. Krug

If one defines *personality* as the totality of influences that explain what a person does in a situation, then personality testing is the process of quantifying those influences. This definition may seem a bit broad at first glance, but the fact is that theorists and researchers have explored a broad range of constructs under the general rubric of personality (see, e.g., McAdams, 1995). Freud, for example, hypothesized a set of internal structures and psychodynamic processes, believing they explained why people react as they do and why they develop certain clinical symptoms. Others have chosen to adopt a more empirical approach to identify and then measure these influences. Raymond B. Cattell, for example, began a programmatic series of empirical research in the 1940s to identify the principal constructs underlying the words people use to describe personality, believing that these underlying dimensions explained observable aspects of personality. Regardless of where the study of personality began, the various lines of thought and research soon led to the proliferation of a variety of instruments designed to quantify key constructs. It also led to the recognition of personality assessment as one of seven specialties and proficiencies in professional psychology by the American Psychological Association (Farberman, 2010).

A vast array of personality tests exists. Some, called *narrow-bandwidth instruments,* were designed to measure a single construct thought to have wide explanatory or predictive value. Others, called *wide-bandwidth instruments,* were built to measure multiple constructs. Some were intended for a single context, for example, clinical assessment or employment selection, and others were thought to measure influences that had predictive utility across contexts. Some never really escaped from the laboratories in which they were developed and the pages of the journal articles in which they were initially described. Others were used by a large body of researchers and users, underwent multiple revisions and improvements, and became very widely used, often in multiple translations across many different languages and cultures.

WHAT MAKES A TEST "OBJECTIVE"?

For many years the term *objective* has been used to distinguish one class of instruments from a second set of tests, such as the Rorschach Inkblot Test (Exner, 1995) and the Thematic Apperception Test (Morgan, 1999), called *projective.* These projective tests use relatively unstructured stimuli such as inkblots and drawings and require a fair degree of sophistication to score. As a result, interscorer reliability became an issue. An alternate format—called *objective*—asked for responses to a standard set of questions, restricted the range of responses, and used straightforward rules to combine the responses into a final score. This format enabled essentially perfect interscorer reliability and more economical testing because the administration and scoring of the test, although not necessarily the interpretation of test scores, required less sophistication and training.

At least in some quarters, the term *objective* has been used to connote a sense of accurate, unbiased,

DOI: 10.1037/14047-019
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

or scientific results provided by one type of instrument versus another, which has been thought of as inaccurate, biased, and unscientific. It should come as no surprise to discover that people who do so are generally not strong supporters of the information derived from the Rorschach or Thematic Apperception Test. To avoid such value terms and their surplus meaning, some writers have suggested the use of the terms *structured* and *unstructured* (Wiggins, 1973). Interestingly, Cattell (1957) reserved the term *objective* specifically for information collected in ways independent of possible distortions with self-evaluations or observer rating (e.g., changes in blood pressure in response to a list of stimulus words).

This chapter limits the domain of objective instruments to those commonly thought of as self-report inventories, checklists, rating scales, and others that use a generally fixed set of questions to which a respondent selects answers from among a generally fixed set of options provided.

ORIGINS OF OBJECTIVE PERSONALITY TESTS

The ancestor of the objective personality test is often thought to be the Personal Data Sheet (Woodworth, 1930), which contained 116 items that represented a variety of anxiety-related symptoms and which was developed in response to a need to screen large populations of army recruits for emotional stability during World War I. It became a model for many more instruments that were subsequently developed to evaluate a variety of clinical symptoms and psychiatric disorders. It also represented a significant departure from the word association tasks and the clinical interview methods that then dominated the evaluation of personality.

Self-report approaches and clinically oriented scales were, however, not the only avenues explored. It is perhaps ironic that some of the earliest work on personality dimensions that was subsequently incorporated into the core of many questionnaire instruments emerged from analyses of peer ratings. R. B. Cattell (1943), Fiske (1949), Norman (1963), and Types and Christal (1961, 1992)—whose work contributed to the definition of the Big Five factors that appear to explain a significant amount of the variance in trait names, at least—all worked, at least initially, with peer ratings. However, because peer ratings restrict data collection to special groups for which a reasonable degree of familiarity can be assumed (e.g., military units, clubs, fraternities), research soon veered toward self-report methods, which were simpler, faster, more economical, and did not impose any special selection requirements on the test population (Ozer & Benet-Martínez, 2006).

METHODS OF TEST DEVELOPMENT

A variety of methods have been used to develop the personality tests that are in widest use today. Although these methods are often described using a variety of terms, it is often simplest to think of them as lying along a continuum ranging from substantive, theoretical considerations at one end to empirical, atheoretical considerations at the other end. In practice, no instrument results from a pure application of one or the other approach, and most result from a blend of development strategies.

A substantive approach is usually guided by an overarching theory or model of personality. The Myers–Briggs Type Indicator (Myers & McCaulley, 1985), which illustrates this approach, was designed to assess personality constructs largely articulated in Jung's theory (Jung & Hull, 1971). That is, the item development and selection process was guided by alignment between the content of the items and the content of the theory.

The advantage of substantive approaches is that theory often provides a clearly stated formulation of the scales and their interaction, and theory can suggest explanations for connections as yet unexplored empirically. The principal disadvantage with the substantive approach is, of course, determining appropriate criteria for evaluating the degree of consistency with the theoretical constructs, which are not directly measurable.

One way of coping with this problem is to explicitly consider a variety of theoretically relevant criteria during test development and validation. Items and scales that yield empirical predictions consistent with theoretical expectations can be regarded as

valid; those that produce inconsistent outcomes can be considered invalid. Such an approach is consonant with modern definitions that consider validity an "integrated evaluation of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989, p. 13). Alternatively, methods of test construction can be used to ensure that scales operate in the same way as the intended constructs. Loevinger (1957, p. 645), for example, discussed the importance of creating personality test scores that are homogeneous with respect to the trait being assessed when the aim of testing is to create instruments of psychological theory rather than simply predictive instruments.

Empirical approaches offer an alternative set of procedures for selecting items and developing test scales. The Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), which is perhaps the most widely used clinical personality inventory, illustrates such an approach, specifically the contrasted-groups approach. Its scales originally represented a collection of diagnostically important symptom clusters rather than a set of theoretical constructs intended to explain clinical symptoms. For example, items were selected for what was originally termed the Paranoia Scale but what has over time more usually been called the Pa Scale or Scale 6 that were able to efficiently differentiate groups of hospital patients diagnosed as paranoid from other patients and asymptomatic samples. There was no requirement that the 40 items that formed the original scale exhibit any degree of internal consistency, for example, or that they correlate in certain ways with other scales, only that they empirically differentiate the two groups.

The obvious advantage of an empirical approach is that the test's validity is not tied to the credibility or viability of a particular theory. Theories and explanations of thought disorders, for example, may change substantially over time, but the scale will still be useful in differentiating people who have delusional beliefs from those who do not.

However, the initial validity of such a scale, at least, rests on the characteristics of the differentiated groups. If changes occur over time in how the groups are conceptualized (i.e., what symptoms constitute a diagnosis), the original item sets may no longer be as useful as they once were. The MMPI–2, which was originally developed in 1939, addressed this problem, although perhaps not consciously, by amassing over time the largest research base of any personality inventory in history. The scales, which themselves underwent a revision, albeit minor, in the 1990s, have been studied with a vast array of populations and have been related to numerous criteria, some involving clinical outcomes, others not. In this way, a great deal of replicated findings related to the test scales has accrued for each of the test scales, thereby substituting a network of empirical findings for a nomothetical network.

RELIABILITY, VALIDITY, AND UTILITY OF PERSONALITY TESTS

From a measurement perspective, a particular strength of objective personality tests is that the reliability of their scores can be directly and easily estimated, and these reliabilities are often shown to be quite high. The use of the plural—*reliabilities*—is intentional because *reliability* is usually defined as the consistency or generalizability of test scores over some change in conditions, and those who use personality tests are usually interested in multiple conditions. Consistency across the items that compose the scale itself is of interest, for example, because the items are usually thought to represent a fixed sampling from a much larger universe of items that could be used to assess the same construct. So the question becomes, for example, "Does the score from this set of dominance items generalize well across the universe of items that reflect dominance, or does it miss critical aspects of self-assertion?"

On this point, personality tests, particularly those developed from a relatively strong theoretical orientation, generally acquit themselves quite well. Many internal consistency coefficients are as high or higher than those found for the most carefully constructed educational achievement tests and reach into the low .90s (Krug, 2004). This may seem trivial if the construct is a relatively narrow one and high internal consistency is achieved by repeating

the same question or very similar questions, which is guaranteed to generate high internal consistency. However, it is not trivial when the construct is relatively broad, such as anxiety, and the universe of content is extensive.

Consistency over time, sometimes called *test–retest reliability* or *score stability,* is also of interest because information collected at one point in time is often used to make decisions (i.e., predictions) at a later point in time. Here, the outcomes are more varied, although most well-crafted personality tests designed to measure stable constructs do show reasonable stability over time, sometimes quite long periods of time (see, e.g., R. B. Cattell, Cattell, & Cattell, 1993; Costa & McCrae, 1992). In this respect, development can be guided by theory, and items designed to measure a transitory or state construct should be expected to show high immediate retest correlations but low long-term correlations.

Validity, the correspondence between empirical evidence and theoretical expectations in Messick's (1989) terminology, is somewhat more complex. As noted earlier, an immense body of empirical evidence exists for some of the personality tests in widest use today: the MMPI–2, 16 Personality Factor Questionnaire (16PF; R. B. Cattell et al., 1993), NEO Personality Inventory—Revised (NEO PI–R; Costa & McCrae, 1992), and Hogan Personality Inventory (HPI; R. Hogan & Hogan, 1995). This research foundation, although extensive, is largely instrument specific. That is, the published articles typically investigate one of these instruments, and it is hard to generalize findings across instruments because the test scales differ.

Research has converged on the conclusion in recent years, however, that many of these instruments, at least those designed primarily to measure "normal range" personality characteristics, exhibit a reasonable degree of content overlap that can be encapsulated in terms of a set of five broad dimensions, which has been termed the *five-factor theory* (Digman, 1990; Wiggins, 1996). The five dimensions are identified as Extraversion (or Surgency), Agreeableness (or Friendliness), Conscientiousness (or Will), Emotional Stability (or Neuroticism), and Intellect (or Openness to Experience or Culture).

The five-factor model has played a fundamental role in guiding the development of newer personality tests such as the NEO PI–3 (McRae & Costa, 2010) and the HPI. Even instruments such as the 16PF that were not originally developed within the context of the five-factor model have attempted to construct a linkage between their underlying personality models and the five-factor model. From a validity perspective, the emergence of the five-factor model has enabled a greater degree of generalization about the validity of these tests. Barrick and Mount (1991), for example, have shown that Conscientiousness as measured by various Big Five instruments showed consistent relations with multiple job performance criteria across a diverse set of occupational groups.

Grucza and Goldberg (2007) conducted what may be the only study in existence of the comparative validity of various personality inventories, which they termed an application of a consumer-testing framework. They examined the predictive power of at least seven tests, some of them the most widely used tests of normal-range personality (e.g., the NEO PI–3, 16PF, HPI, and California Psychological Inventory [Gough & Bradley, 2002]) with respect to six clusters of behavioral acts about which respondents provided frequency information. Two of the clusters addressed acts generally regarded as undesirable (drug use, undependability), and two addressed desirable acts (friendship, creative achievement). The final two were neutral in their desirability (reading, writing). A sample act would be, for example, "started a conversation with a stranger," for which the answers ranged from *never in my life* to *more than 15 times in the past year*. Although the degree of predictability across all the instruments was generally high, the best in the study (16PF) was able, on average, to predict almost 10% more of the criterion variance than the lowest (Temperament and Character Inventory; Cloninger, Przybeck, Svrakic, & Wetzel, 1994).

Criticisms occasionally arise that self-report measures are of little practical use in predicting behavior and that observer ratings (i.e., peer ratings), which seem to have gained very wide acceptance in organizational settings in recent years, are themselves subject to considerable distortion. Some of the reasons

offered in support of such assertions are that stylistic factors, such as social desirability, play a greater role in answer selection than content factors or that self-report inventories are too transparent and sensitive to deliberate distortion to provide valid information. A job applicant, for example, particularly one who was out of work, might be hard pressed to be entirely self-revelatory.

Of course, all assessment, not just personality tests is subject to distortion (Krug & Cattell, 1971). Anyone involved in large-scale educational assessment for any time, for example, has found the occasional answer sheet on which the responses form a picture or some interesting pattern. Perhaps the concerns about the accuracy of objective personality tests arise simply from the fact that they are among the most highly researched and critically examined of instruments. Quality test design considers various important sources of variance that may influence responses in the selection of items, not just social desirability or faking-good tendencies. Then the items and scales are structured to maximize substantive variance and minimize irrelevant factors. The Jackson Personality Inventory (Jackson, 1994) provides an excellent example of this approach to test construction.

With respect to deliberate faking, R. Hogan and Hogan (1995) have argued that the ability to alter scores on personality measures is a function of social competence and valuable information in itself. Moreover, they pointed out that the base rate for faking in the job application process is much lower than might have been thought. Perhaps this explains why personality testing enjoys such widespread use in employment settings.

There are, however, some areas in which a degree of caution should be taken, and always is not: the use of personality tests with special populations. For example, literacy limitations in the population with hearing impairment are often significant, and the use of tests designed for hearing populations may produce misleading results (Rosen, 1967). Culture represents an important contextual variable as well. Eyde, Robertson, and Krug (2010) reported a case in which the use of the MMPI–2 in English with Chinese nationals, who had only a working knowledge of the language, produced distorted results that could be explained on the basis of researched and documented cultural differences. The deceptive simplicity of objective personality tests has led, on occasion, to their being used in inappropriate contexts with misleading results.

AREAS OF APPLICATION

Objective personality tests have been and continue to be used in a variety of contexts. Clinical assessment, of course, remains a major application area, but objective personality tests enjoy widespread use in industrial/organizational, forensic, and school contexts as well.

Clinical Assessment

It goes without saying that objective personality tests are widely used in clinical settings. This area is probably the one in which they enjoy the greatest use and the greatest utility, as least in the opinion of some.

Clinically oriented instruments such as the MMPI–2, the Millon Clinical Multiaxial Inventory—III (MCMI–III; Jankowski, 2002), or the Personality Assessment Inventory (Morey, 2007) provide information that is useful in diagnosing psychopathology and personality disorders. Clinical instruments can be very helpful in documenting the degree of initial impairment and improvement during therapy.

Instruments oriented to normal-range personality characteristics have, however, been widely used in clinical settings as well. Beyond the information provided by clinical scales, information about characteristics such as Conscientiousness and Openness to Experience, two of the Big Five, has proven to be very helpful in predicting the course of therapy and structuring reasonable treatment plans. Cloninger (1987), for example, described two alcoholism syndromes whose personality dynamics are very different. One type is characterized by anxiety accompanied by feelings of guilt and shame, and the other type is characterized by antisocial personality characteristics, with impulsive consumption of alcohol that is often accompanied by fighting and subsequent arrest. Other individual-differences characteristics are helpful in anticipating who will have greater difficulty following through on a treatment plan and who might need more support.

Industrial and Organizational Assessment

Personality testing in work settings, which was well established by 1970, endured a period of exile during the 1970s and 1980s, not because of increased criticism of its utility or predictability but because of concerns about the legality of such testing in light of changes in federal legislation. However, since the 1990s it has reemerged even stronger than before.

As noted earlier, the use of personality tests for large-scale assessment began during World War I and arose from the need to screen large populations economically and quickly. During World War II, Henry Murray headed a team of personality researchers who pioneered innovative assessment methods—some objective tests, some performance tests, some projective instruments—for assessing leadership characteristics, resiliency, and other characteristics that industry found useful in evaluating potential executives and managers.

Barrick and Mount's (1991) research showed that Conscientiousness, a dimension reliably assessed by many different personality tests, predicted multiple job performance criteria not only for professionals and managers, but also for police and those in skilled and semiskilled jobs. Extraversion and Openness to Experience both predicted success in training across all the groups. For reasons that relate ultimately to efficiency and utility, personality tests are very widely used in organizational settings today. The accessibility of the Internet and the structure of objective personality tests have enabled them to be conveniently and widely administered at a distance, and many companies use this feature to conduct an initial screening of applicants, particularly those for sales positions and those requiring interpersonal skills.

The connection between personality characteristics and job success has also made it possible for personality tests to be widely used in career counseling and exploration. Although a distinction is often made between personality tests and career inventories, the line between them is blurry (Krug, 1995), and it is hard to know exactly, for example, whether Holland's Vocational Preference Inventory should be thought of as a career inventory or a personality test, particularly because some of the initial evidence for the construct validity of the Vocational Preference Inventory stemmed from its relationship to the 16PF (Holland, 1960).

Other Areas

Forensic assessment, which may be thought of as a specific subarea within clinical assessment although its practitioners see it as distinct, uses personality tests extensively and has experienced substantial growth in recent years. Membership in Division 41 (American Psychology-Law Society) of the American Psychological Association increased to 2,046 in 2006 from 1,153 in 1987 (Eyde et al., 2010). Tests such as the MMPI–2 have become almost standard in custody evaluations, for example, or when conducting evaluations of culpability and responsibility.

School psychologists make use of personality tests in educational settings, especially to understand learning or adjustment problems, although this arena requires perhaps an even greater degree of sensitivity to ethical and policy issues (Knauss, 2001).

Although it might be considered within the scope of clinical or industrial testing, personality testing enjoys widespread use in the area of law enforcement selection. A survey of practice in this area (Super, 2006) found that more than half of the 478 federal, state, and local law enforcement agencies used the MMPI–2, the Inwald Personality Inventory (Inwald, 2008), or the California Personality Inventory (Gough & Bradley, 2002).

MAJOR INSTRUMENTS

This section provides a brief overview of some of the objective personality tests in widest use today: MMPI–2, MCMI–III, 16PF, NEO PI–3, and HPI. The first two are oriented principally to the realm of pathological personality, and the last three are oriented to the realm of normal personality characteristics. All have extensive research foundations, and several, at least, have been translated, adapted, and standardized in other countries.

Minnesota Multiphasic Personality Inventory—2

The MMPI–2 is the 1989 revision of a test that first appeared in 1939. As noted earlier, test construction

relied on the contrasted groups or empirical keying approach, and item selection was determined by each item's ability to differentiate clinically diagnosed groups of adults. Most adults with normal reading skills finish the test in an hour or less.

The test was originally designed around a set of 10 clinical scales. Over time, and with additional research contributed by many different authors, the number of scales vastly increased, although the clinical scales probably remain the core of the test for most users. Identified initially by labels such as Hypochondriasis, Paranoia, and Psychopathic Deviate, it became fairly standard over time to refer to the scales by number (e.g., Scale 1, Scale 6, Scale 4), thereby avoiding the psychiatric nomenclature of the 1930s and 1940s. Four validity scales, Cannot Say, Lie (L), Infrequency (F), and Defensiveness (K), became an integral part of the basic test profile.

A radical departure from the original instrument came with the introduction of the revised clinical scales (Tellegen et al., 2003), which attempted to reduce overlap among the original clinical scales. The revised clinical scales were developed after a careful psychometric strategy and offered as a more refined version of the original clinical scales. Their introduction, however, has not been universally applauded (Caldwell, 2006; Nichols, 2006). However, the authors of the revised clinical scales have made an impressive argument for their utility, and there is some reason to believe they might eventually replace the original clinical scales. To facilitate that process, a new version of the test, the MMPI–2—Revised Form (Ben-Porath & Tellegen, 2008), can be scored only for the revised clinical scales, not for the original clinical scales.

The MMPI–2 is probably the most extensively researched personality test in the world. Thousands of articles have been published, a great many by independent researchers, that examine the predictability of the test scales in a wide variety of settings, across cultures, languages, and time. The test is not without its many critics, although the revisions have attempted to address some of the most important psychometric qualities of the basic scales (e.g., Tellegen et al., 2003) and the representativeness of the normative sample. Nonetheless, the test's wide use attests to its utility and acceptance by a vast array of users.

Millon Clinical Multiaxial Inventory—III

As with the MMPI–2, the MCMI–III was developed for clinical use within a clinical population. The first edition appeared more than 30 years ago; the current edition appeared in 2009. The test is shorter than the MMPI–2, and it generally takes 20 to 30 minutes for the average patient to complete the 175 questions.

Another point of departure from the MMPI–2 is that a theoretical model of psychopathology explicitly guided test development. The scales were intended to represent diagnostic categories described in the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000). Items were initially written to correspond directly to symptom descriptions. A set of analyses then proceeded to eliminate items that correlated substantially with other scales and retain those that showed a high degree of internal consistency. Finally, the items and scales were evaluated in terms of their ability to distinguish diagnostic groups. Rather than report scores in terms of a set of normalized standard scores, as most personality tests do, the MCMI–III uses base-rate scales, which attempt to take into account the differing prevalence rates of the diagnostic category associated with each scale.

Although its bibliography is not as rich as that of the MMPI–2, the MCMI–III has amassed in excess of 500 research articles, a very respectable showing by any criterion, particularly considering that the only article written about many personality tests is the one in which they were first described.

16 Personality Factor Questionnaire

The 16PF is one of the most widely used, theory-based instruments for assessing normal-range personality characteristics in adults. Since its first U.S. publication in 1949, the test has been translated into nearly 50 languages. The test is used worldwide to evaluate a set of 16 reasonably independent personality characteristics that predict a wide range of socially significant criteria.

Raymond B. Cattell and a series of coauthors developed the test over many decades on the basis of extensive research intended to clarify the basic organization of human personality. Cattell was

interested primarily in identifying a relatively small set of "source traits" that could be used to explain variations in the much larger set of "surface" characteristics observable in behavior and recorded in language. Cattell looked to language in his search because he was convinced that "all aspects of human personality which are or have been of importance, interest, or utility have already become recorded in the substance of language" (R. B. Cattell, 1943, p. 478). His starting point was the work of Allport and Odbert (1936), who had identified about 18,000 words in an English-language dictionary that described distinctive aspects of human behavior. When they eliminated terms that were essentially evaluative (e.g., *adorable, evil*), were metaphorical (e.g., *alive, prolific*), or described temporary states (e.g., *rejoicing, frantic*), 4,504 terms still remained. Cattell conducted a series of analyses on this lexicon to eliminate overlap among them. The first publication of the test did not occur until 1949, more than a decade after these studies began. Since then, the test has undergone several major, and more numerous minor, revisions. The most recent, in 1993, was the last Cattell completed before his death in 1998.

The primary scales of the test, which are designated by alphanumeric symbols, are as follows: A, Warmth; B, Reasoning; C, Emotional Stability; E, Dominance; F, Liveliness; G, Rule-Consciousness; H, Social Boldness; I, Sensitivity; L, Vigilance; M, Abstractedness; N, Privateness; O, Apprehension; Q_1, Openness to Change; Q_2, Self-Reliance; Q_3, Perfectionism; and Q_4, Tension. Five global factors (Extraversion, Anxiety, Tough-Mindedness, Independence, and Self-Control) assess features similar to those defined by the five-factor model. Besides the primary scales and global factors, the 16PF is scored for approximately 100 derivative scales (e.g., Creativity, Leadership), criteria that derive from years of research on 16PF applications in clinical, counseling, and organizational psychology. The 16PF also provides three response style indicators: Impression Management, Infrequency, and Acquiescence. These scales are helpful in identifying unusual response patterns that may affect the validity of the profile.

The 16PF is a challenging instrument with which to work. Much of Cattell's extensive research on the instrument was directed toward theoretical and psychometric concerns. He paid less attention to practical issues of profile analysis or clinical interpretation. Fortunately, other authors have developed a variety of interpretive resources that help users understand the meaning of the scales in a variety of contexts (H. B. Cattell, 1989; Karson, Karson, & O'Dell, 1997; Krug, 1981; Lord, 1997, 1999).

The 16PF is an important instrument whose utility has been enhanced by extensive research and by periodic updating. It is theoretically grounded in research on the basic structure of adult personality and represents a significant resource for decision makers in a variety of settings. Despite its age, it continues to attract significant research interest, which has, in fact, been growing. An examination of 3,127 research references to the test found in a PsycINFO search showed that more than a third had been published in the past 5 years.

NEO Personality Inventory—3

The NEO Personality Inventory, most recently revised in 2010 (the NEO PI–3; McCrae & Costa, 2010), was specifically developed to assess the five-factor model. *NEO* derives from Neuroticism, Extraversion, and Openness, for which established facet scales existed when the test first appeared (Costa & McCrae, 1985). The 1992 revision (NEO PI–R; Costa & McCrae, 1992) added facet scales for Agreeableness and Conscientiousness, which were assessed only globally in the first edition.

Two versions of the test exist, one for self-report (Form S) and another for external ratings (Form R). This feature is interesting and unusual among personality tests, which do not typically provide a parallel instrument for collecting information from outside raters. The test intentionally does not incorporate validity scales per se because the authors believe they may actually detract from the validity of the instrument (McCrae et al., 1989). However, Costa and McCrae (1992) presented a series of indicators derived from the performance of a large volunteer sample that provide a check on some common response styles. In addition, the last three questions of the test ask the test taker directly whether he or she has honestly and accurately answered all of the questions and entered the

responses correctly. Except, perhaps, for those provided by the out-of-work job applicant or psychopath, answers to these questions may be as informative as scores on more sophisticated validity scales.

In the 20 years or so since it was first published, an extensive library of research findings has been published (Costa & McCrae, 2003). In addition, given the ubiquity of the five-factor model in contemporary personality research, the NEO PI–R is well poised to benefit from the accumulation of findings generated by interest in the model itself.

Hogan Personality Inventory

The HPI was designed primarily for use in personnel selection, employee development, and career-related decision making. It assesses characteristics that aid in understanding how people get along with others and how they achieve educational and career goals. The entire test can be completed in 15 to 20 minutes and requires about a fourth-grade reading level.

R. Hogan and Hogan (1995) relied explicitly on the five-factor model in the test development process but adjusted the final form to the structure suggested by their own analyses. HPI development relied extensively on samples of employed adults, whereas most other tests have made heavy use of college student respondents. Although personality test developers over the past half-century or more have operated on the assumption that the two populations are interchangeable, there probably are some differences, and the HPI benefits from actually having been developed on the population with which it is intended to be used.

The test manual (R. Hogan & Hogan, 1995) provides an interesting summary of validity data that relates HPI scales to peer descriptions and to various aspects of organizational behavior. The former is important to the validity of the test because a primary goal of the authors was to create an instrument that predicted how others would describe a test taker. The latter is refreshing because personality tests are regularly applied, often uncritically, to predict job performance that sounds like the name of a test scale (i.e., employee theft from Integrity, promotion potential from Ambition). It often works, but it is nevertheless comforting to see that the test authors can provide documentation, not simply assertions. Meta-analyses of HPI validity data are available in the professional literature (J. Hogan & Holland, 2003).

The HPI is a relative newcomer compared with tests such as the MMPI–2 or the 16PF. However, since its introduction at the 1982 Nebraska Symposium, its use has expanded rapidly. The carefully developed validity data that have accumulated so far would seem to predict even wider use for it in years to come.

ETHICAL ISSUES IN THE USE OF OBJECTIVE PERSONALITY TESTS

The deceptive simplicity of objective personality tests presents some ethical concerns. Two in particular are the level of training required of test users and the range of application for a single instrument.

Because objective personality tests are usually very straightforward to administer and score, the task can be, and too often is, handed off to someone who has little familiarity with testing at all. Eyde et al. (2010) reported a case in which personality testing for a college scholarship was handled by the school's nursing staff after a brief orientation provided by a psychologist, who also furnished the staff with a set of decision rules to apply to MMPI–2 results in determining an accept–reject decision. In this case, ease of administration and scoring led to a belief that interpretation was easy, too. Without proper training, however, test users may reach unsupportable conclusions and incorrect decisions.

Because objective tests can usually be completed with little or no interaction required between the test taker and the administrator, situations exist in which test takers are given the materials to take home and return at the next visit. In hiring situations, the ease with which objective personality tests can be presented online tempts those in charge to have candidates complete the tests before they appear for an interview. In either case, the accuracy of the results should be in considerable doubt.

Although it may not take a rocket scientist to administer or score the test, it does take a reasonable level of scientific knowledge to reach valid conclusions on the basis of test results. In the final analysis,

many criticisms of objective tests may be directed not toward the intrinsic features of those tests but toward misuses of test results by inadequately qualified users.

A second concern is that the ease-of-use feature of objective personality tests might encourage users to implement them in settings or for purposes for which they were never intended. Just because a test can be used in a situation does not mean that it should be. The MMPI–2 was originally designed to provide diagnostic information primarily within a clinical population. It has become very widely used, however, for purposes the authors probably never envisioned. The 16PF, though, is not well suited to assess major affective and cognitive disturbances. Experienced interpreters of the profile have identified certain score patterns suggestive of depression (Karson et al., 1997) and argued for its clinical interpretation, but diagnosis from the 16PF scales alone is usually difficult or impossible for the vast majority of those who work with the test.

In each situation in which objective personality tests are used, a conscious, deliberate evaluation should be made of its appropriateness. If it seems a stretch of purpose, it probably is, and the application should be rethought.

CURRENT DEVELOPMENTS IN PERSONALITY ASSESSMENT

Recent developments in measurement theory and computerization have had an impact on the way in which personality tests are developed and used. Efforts have also been directed toward the development of nonproprietary item pools, which can be used to assess a variety of important personality characteristics that were previously only measurable through proprietary instruments.

Developments in Measurement Models

Most personality assessment instruments in wide use were developed within the context of classical test theory. This was also true, of course, for cognitive assessment, but in the course of the past several decades, cognitive assessment has energetically embraced an alternative set of models, which are collectively referred to as item response theory (IRT). Whereas classical test theory can be thought of as a theory of test performance, IRT can be thought of as a theory of item performance. More specifically, IRT explains an individual's response to a test item in terms of the person's trait or proficiency level interacting with various characteristics of items, such as difficulty and discrimination. Multiple formulations of this relationship exist. One of the most commonly encountered is

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + \exp\left[-Da_j(\theta_i - b_j)\right]}.$$

In the cognitive domain in which IRT models have been most extensively applied, $P_j(\theta_i)$ represents the probability of a correct response, a_j represents the item discrimination, b_j the item difficulty, and c_j the probability of a correct response by a very low-scoring individual, sometimes called the *pseudo-guessing parameter*. D is a scaling factor that brings the interpretation of the logistic model parameters in line with those of the normal ogive model. When IRT models are extended to personality scales, some shift in terminology is required because personality scales do not have correct answers, and examinees do not usually guess the answer. Instead, it is probably reasonable to interpret $P_j(\theta_i)$ as the probability of endorsement, b as the location of the item on the underlying dimension (i.e., does the item represent a relatively higher or lower trait level?), and c as the probability of endorsement by an individual who scores very low on the dimension. Simplifying assumptions about the nature of the test items makes it possible to eliminate certain parameters, thus simplifying the model, and terms have been added to handle items that have more than one single answer (i.e., Likert-type response formats), which are very common in personality assessment, rather than a dichotomous, yes–no response format (e.g., Bock, 1972; Muraki, 1992).

Research has begun to accumulate evidence for the utility of IRT models in personality assessment (Reise & Henson, 2003). One of the advantages of using IRT models is that item parameter values are independent of the samples used to estimate them when the assumptions of the model are met. If a pool of items is calibrated to a common standard,

then different sets of items derived from the pool (i.e., test forms) will produce a score on the common scale without the need to conduct separate normative studies for each test form. This feature—which is highly desirable in applications such as credentialing, in which test forms must be changed continuously and often frequently to avoid overexposure of specific item content—may be less valuable in personality assessment, in which test forms tend to remain stable for years or decades without any noticeable deterioration in the usefulness of the test. However, personality testing has usually ignored the issue of item exposure. The same set of items is often administered repeatedly to evaluate change, for example, without considering the impact of memory for specific item content on the score.

Perhaps a more valuable feature is that IRT allows items to be selected for administration that provide the highest degree of precision in targeted regions of the trait distribution. That is, IRT allows users to focus the instrument's bandwidth at a particular point on the scale. This means, for example, that items to measure improvement within a hospitalized population might be selected that represent relatively higher levels of the underlying dimension. Traditional, fixed-level personality tests usually include a quasi-normal distribution of item locations to be useful with the widest possible population of test takers.

Computerization of Personality Assessment

Personality assessment recognized the advantages of computer use early. Within a decade after electronic computers first became available for general use, and long before the advent of personal computers, the first computer interpretive reports of personality tests appeared. The MMPI report developed at the Mayo clinic (Swenson, Rome, Pearson, & Brannick, 1965) in 1962 may have been the first, but it was certainly not the last. In the next quarter century, hundreds of computer-based test interpretation systems followed (Butcher, Perry, & Atlis, 2000; Krug, 1993).

These systems often began simply as scoring aids, automating the process of aggregating item responses to form scales. They soon, however, began to address the issue of profile interpretation as well. For some extensively researched tests such as the MMPI–2 there were vast libraries of interpretive hypotheses about individual scale elevations and configural patterns that were difficult for unassisted human interpreters to keep in mind. In this case, the computer supplemented human memory and allowed systems to generate reports with a richer set of interpretive possibilities than might otherwise have been possible, at least in the computer-based interpretation time frame.

One of the more interesting recent developments in the area of computer applications to personality assessment is the Patient-Reported Outcomes Measurement Information System (PROMIS), a collaborative project funded by the National Institutes of Health to create a set of nonproprietary, IRT-calibrated item pools for assessing physical and mental health outcomes from the patient's perspective. PROMIS also offers a computer-adaptive testing system to facilitate administration of the item pools through the Internet (Cella et al., 2007, 2010). Because the focus is on health status assessment, the item pools are necessarily oriented toward constructs related to mental and social health (e.g., anxiety, depression, anger, social isolation). The availability of a tool for online testing may represent a significant enhancement in personality research in particular.

The International Personality Item Project (Goldberg et al., 2006) is another development that has significant potential for shaping the future of personality assessment. The International Personality Item Project began from the perspective that, although most narrow-bandwidth personality inventories are in the public domain and can be freely used and researched by scientists around the world, broad-bandwidth inventories are proprietary. Although the constructs they measure are of potentially greater interest to personality researchers, others cannot freely and easily use these items (Goldberg, 1999). Public domain item sets have been developed for most of the major multidimensional personality inventories (e.g., NEO PI, 16PF, HPI) as well as scales to measure the Big Five and constructs such as emotional intelligence. Although scales exist for depression, anxiety, and dissociation, scales for the clinical realm are

relatively less represented in the bank of 2,413 items, which may represent only a current limitation of the International Personality Item Project because many of the items (e.g., "Can get anxious, depressed, or irritable for no reason," "Believe that people are essentially evil," "Believe that I have a serious disease," "Begin to panic when there is danger") cover the same kinds of content represented in inventories such as the MMPI–2 and other multipurpose clinical inventories.

Issues relating to cross-cultural aspects of personality testing are likely to become only more prevalent in the future. The population of the United States has become increasingly diverse, and as it does, the need for instruments or adaptations of instruments that correctly represent the underlying constructs across language and cultural differences becomes more urgent. Many of the major personality instruments have already addressed the language issue. Translations and adaptations exist in many languages for the 16PF, the MMPI–2, and other well-researched instruments, and portions of the International Personality Item Project item pool have been translated into several dozen languages.

Such translations are an important first step in ensuring that the instruments are appropriate for use with cultural minority populations, including those in the United States. As illustrated in Eyde et al.'s (2010) case study regarding the use of the MMPI in translation, translation is not the final step in the process of ensuring culturally neutral instruments. Differences may extend to norms, administration procedures, and other elements of the testing process (Geisinger, 1994, 1998). As the population of potential personality test takers becomes increasingly diverse, the need for better instruments and the research to document their utility will only increase.

CONCLUSION

Objective personality tests provide useful information that contributes to the effectiveness of decisions about people. Existing instruments incorporate extensive research bases that address a wide variety of relevant issues. Newer instruments, revisions of well-established instruments, and newer approaches to personality assessment itself represent increasingly sophisticated products of psychometric knowledge.

References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*, i–171. doi:10.1037/h0093360

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Barrick, M., & Mount, M. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

Ben-Porath, Y. S., & Tellegen, A. (2008). *Minnesota Multiphasic Personality Inventory—2—Revised Form*. Minneapolis: University of Minnesota Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51. doi:10.1007/BF02291411

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality—2 (MMPI–2): Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Butcher, J. N., Perry, J. N., & Atlis, M. M. (2000). Validity and utility of computer-based test interpretation. *Psychological Assessment, 12*, 6–18. doi:10.1037/1040-3590.12.1.6

Caldwell, A. B. (2006). Maximal measurement or meaningful measurement: The interpretive challenges of the MMPI–2 Restructured Clinical (RC) scales. *Journal of Personality Assessment, 87*, 193–201. doi:10.1207/s15327752jpa8702_09

Cattell, H. B. (1989). *The 16PF: Personality in depth*. Champaign, IL: Institute for Personality and Ability Testing.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38*, 476–506. doi:10.1037/h0054116

Cattell, R. B. (1957). *Personality and motivation: Structure and measurement*. New York, NY: Harcourt, Brace, World.

Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993). *16PF Fifth Edition Questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Lai, J.-S. (2010). The Patient Reported Outcomes Measurement Information System

(PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*, 1179–1194. doi:10.1016/j.jclinepi.2010.04.011

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Matthias, R. (2007). The Patient Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group during its first two years. *Medical Care*, *45*(5, Suppl. 1), S3–S11. doi:10.1097/01.mlr.0000258615.42478.55

Cloninger, C. R. (1987). Neurogenetic adaptive mechanisms in alcoholism. *Science, 236*, 410–416. doi:10.1126/science.2882604

Cloninger, C. R., Przybeck, T. R., Svrakic, D. M., & Wetzel, R. D. (1994). *The Temperament and Character Inventory (TCI): A guide to its development and use*. St. Louis, MO: Center for Psychobiology of Personality, Washington University.

Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI–R) and NEO Five-Factor Inventory (NEO FFI) manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (2003). *Bibliography for the Revised NEO Personality Inventory (NEO PI–R) and NEO Five Factor Inventory (NEO FFI)*. Lutz, FL: Psychological Assessment Resources.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440. doi:10.1146/annurev.ps.41.020190.002221

Exner, J. E. (1995). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations*. New York, NY: Wiley.

Eyde, L. D., Robertson, G. J., & Krug, S. E. (2010). *Responsible test use: Case studies for assessing human behavior* (2nd ed.). Washington, DC: American Psychological Association.

Farberman, R. K. (2010). Council recognizes seven specialties and proficiencies. *Monitor on Psychology, 41*(9), 74–75.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44*, 329–344. doi:10.1037/h0057198

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312. doi:10.1037/1040-3590.6.4.304

Geisinger, K. F. (1998). *Psychological testing of Hispanics*. Washington, DC: American Psychological Association.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96. doi:10.1016/j.jrp.2005.08.007

Gough, H. G., & Bradley, P. (2002). *The California Psychological Inventory manual* (3rd ed.). Palo Alto, CA: CPP, Inc.

Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment, 89*, 167–187. doi:10.1080/00223890701468568

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100–112. doi:10.1037/0021-9010.88.1.100

Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory manual* (2nd ed.). Tulsa, OK: Hogan Assessment Systems.

Holland, J. L. (1960). The relation of the Vocational Preference Inventory to the Sixteen Personality Factor Questionnaire. *Journal of Applied Psychology, 44*, 291–296. doi:10.1037/h0049367

Inwald, R. (2008). The Inwald Personality Inventory (IPI) and Hilson Research Inventories: Development and rationale. *Aggression and Violent Behavior, 13*, 298–327. doi:10.1016/j.avb.2008.04.006

Jackson, D. N. (1994). *Jackson Personality Inventory manual revised*. Port Huron, MI: Sigma Assessment Systems.

Jankowski, D. (2002). *A beginner's guide to the MCMI–III*. Washington, DC: American Psychological Association. doi:10.1037/10446-000

Jung, C. G., & Hull, R. F. C. (1971). Psychological types. In H. Read, M. Fordham, G. Adler, & W. McGuire (Eds.), *Collected works of C. G. Jung* (Vol. 6, pp. 510–523; H. G. Baynes, Trans.). Princeton, NJ: Princeton University Press.

Karson, M., Karson, S., & O'Dell, J. (1997). *16PF interpretation in clinical practice: A guide to the fifth edition*. Champaign, IL: Institute for Personality and Ability Testing.

Knauss, L. K. (2001). Ethical issues in psychological assessment in school settings. *Journal of Personality Assessment, 77*, 231–241. doi:10.1207/S15327752JPA7702_06

Krug, S. E. (1981). *Interpreting 16PF profile patterns.* Champaign, IL: Institute for Personality and Ability Testing.

Krug, S. E. (1993). *Psychware: A reference guide to computer-based products for assessment in psychology, education, and business.* Champaign, IL: MetriTech.

Krug, S. E. (1995). Career assessment and the Adult Personality Inventory. *Journal of Career Assessment, 3*, 176–187. doi:10.1177/106907279500300205

Krug, S. E. (2004). The Adult Personality Inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (3rd ed., Vol. 3, pp. 679–694). Mahwah, NJ: Erlbaum.

Krug, S. E., & Cattell, R. B. (1971). A test of the trait-view theory of distortion in measurement of personality by questionnaire. *Educational and Psychological Measurement, 31*, 721–734. doi:10.1177/001316447103100312

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Lord, W. (1997). *16PF5: Personality in practice.* Windsor, England: NFER-Nelson.

Lord, W. (1999). *16PF5: Overcoming obstacles to interpretation.* Windsor, England: NFER-Nelson.

McAdams, D. P. (1995). What do we know when we know a person? *Journal of Personality, 63*, 365–396. doi:10.1111/j.1467-6494.1995.tb00500.x

McCrae, R. R., & Costa, P. T. (2010). *Professional manual for the NEO Inventories.* Lutz, FL: Psychological Assessment Resources.

McCrae, R. R., Costa, P. T., Jr., Dalhstrom, W. G., Barefoot, J. C., Siegler, I. C., & Williams, R. B., Jr. (1989). A caution on the use of the MMPI K-correction in research on psychosomatic medicine. *Psychosomatic Bulletin*, *51*, 58–65.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Phoenix, AZ: Oryx Press.

Morey, L. C. (2007). *The Personality Assessment Inventory professional manual.* Lutz, FL: Psychological Assessment Resources.

Morgan, W. G. (1999). The 1943 TAT images: Their origin and history. In M. T. Gieser & M. I. Stein (Eds.), *Evocative images: The Thematic Apperception Test and the art of projection* (pp. 65–83). Washington, DC: American Psychological Association. doi:10.1037/10334-005

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. doi:10.1177/014662169201600206

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers–Briggs Type Indicator.* Mountain View, CA: CPP, Inc.

Nichols, D. S. (2006). The trials of separating bath water from baby: A review and critique of the MMPI–2 Restructured Clinical scales. *Journal of Personality Assessment, 87*, 121–138. doi:10.1207/s15327752jpa8702_02

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*, 574–583. doi:10.1037/h0040291

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*, 401–421. doi:10.1146/annurev.psych.57.102904.190127

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93–103. doi:10.1207/S15327752JPA8102_01

Rosen, A. (1967). Limitations of personality inventories for assessment of deaf children and adults as illustrated by research with the Minnesota Multiphasic Personality Inventory. *Journal of Rehabilitation of the Deaf, 1*, 47–52.

Super, J. T. (2006). A survey of pre-employment psychological evaluation tests and procedures. *Journal of Police and Criminal Psychology, 21*, 83–87. doi:10.1007/BF02855686

Swenson, W. M., Rome, H. P., Pearson, J. S., & Brannick, T. L. (1965). A totally automated psychological test experience in a medical center. *JAMA, 191*, 925–927. doi:10.1001/jama.1965.03080110049012

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI–2 Restructured Clinical Scales: Development, validation, and interpretation.* Minneapolis: University of Minnesota Press.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Wiggins, J. S. (1996). *The five-factor model of personality: Theoretical perspectives.* New York, NY: Guilford Press.

Woodworth, R. S. (1930). Autobiography of Robert S. Woodworth. In C. A. Murchison (Ed.), *History of psychology in autobiography* (Vol. 2, pp. 359–380). Worcester, MA: Clark University Press.

CHAPTER 20

PERFORMANCE ASSESSMENT IN EDUCATION

Suzanne Lane

Performance assessments are considered by policymakers and educators to be valuable tools for educational reform (Linn, 1993; Resnick & Resnick, 1992). Performance assessments that measure students' thinking and reasoning skills and their ability to apply knowledge to solve meaningful problems will help shape sound instructional practices by modeling to teachers what is important to teach and to students what is important to learn (Lane, 2010). They serve as exemplars that stimulate and enrich learning rather than just serve as indicators of learning (Bennett & Gitomer, 2009); they are contextualized, linking school activities to real-world experiences (Darling-Hammond, Ancess, & Falk, 1995); and they can include opportunities for self-reflection and collaboration as well as student choice, such as choosing a particular topic for a writing assessment (Baker, O'Neil, & Linn, 1993). Performance assessments attempt to "emulate the context or conditions in which the intended knowledge or skills are actually applied" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Medical Education [NCME], 1999, p. 137). They may range from asking students to provide a rationale for their mathematics problem to conducting a scientific investigation. This chapter focuses on the status and uses, design and scoring, and validity of performance assessments.

STATUS AND USES OF PERFORMANCE ASSESSMENTS

Unfortunately, the use of performance assessments in large-scale assessment systems declined with the requirements of the No Child Left Behind Act of 2001 (U.S. Department of Education, 2005). With the advent of the Common Core State Standards initiative (Council of Chief State School Officers & National Governors Association, 2010) and the U.S. Department of Education Race to the Top initiative (U.S. Department of Education, 2009), there is a renewed interest in using performance assessments in state assessment systems that are grounded in academic standards that reflect 21st-century skills. The Common Core State Standards represent a set of expectations for knowledge and skills that students need to be prepared for success in college and careers when they graduate from high school. The Common Core State Standards emphasize students' ability to reason, synthesize information from various sources, think critically, and solve challenging problems. Most states have adopted the Common Core State Standards and have competed for federal funding from the U.S. Department of Education's (2009) Race to the Top initiative. A major goal of the Common Core State Standards and the Race to the Top initiative is to help ensure that academic standards are set high for all students. To be awarded federal funding, states need to show how their assessment system will measure "standards against which student achievement has traditionally been difficult to measure" and include "items that will be varied and elicit complex student demonstrations or applications of knowledge and skills" (U.S. Department of Education, 2010, p. 8).

In response to these initiatives, the proposed Balanced Assessment System calls for periodic through-course performance tasks and an end-of-year

DOI: 10.1037/14047-020
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

reference exam that are used for summative, formative, accountability purposes (Darling-Hammond & Pecheone, 2010). The performance tasks are to be curriculum embedded (e.g., exhibitions, product developments), standardized, and administered and scored by teachers. The end-of-year reference exams are intended to include various item formats (e.g., selected response, short and extended answer, complex electronic items), be computer adaptive, be scaled vertically across a range of learning progressions, and use both computer-automated scoring and moderated human scoring. Although several design and psychometric issues need to be addressed for this new generation of assessment systems, we need to embrace this opportunity so as to ensure that assessments not only include performance tasks that assess complex thinking skills but are also positioned to inform and enhance both teaching and student learning.

Performance assessments can be used for formative assessments, interim assessments, and summative assessments. Formative assessments are embedded within curriculum and instruction to promote student learning and are used by teachers to diagnose where students are in their learning, identify gaps in student understanding, and promote student learning. Interim or periodic assessments are administered by the school or district several times a year, the results can be meaningfully aggregated to the school or district level (Perie, Marion, & Gong, 2009), and they can inform decisions not only at the student level, but also at the class, school, and district levels. Some of their purposes include predicting student scores on large-scale assessments, evaluating an educational program, diagnosing gaps in student learning, or providing information for an accountability program. Summative assessments can be used at the end of the school year or term to assess student performance against content standards. As an example, state summative assessments are used as part of an accountability program to inform policy.

DESIGN AND SCORING OF PERFORMANCE ASSESSMENTS

The design of performance assessments begins with the delineation of the conceptual framework, including a description of the purpose of the assessment, the concept to be assessed, and the intended inferences to be drawn from the assessment results (Lane & Stone, 2006). With regard to the type of inferences, one may want to generalize to the larger construct domain of interest or provide evidence of a particular accomplishment or performance. The former approach entails sampling tasks from the entire targeted domain to ensure content representativeness, and the latter approach entails specifying a performance task that allows for the demonstration of a broader ability or performance such as a high school project.

The extent to which the design of a performance assessment considers theories of student learning will affect the validity of score inferences. Therefore, the conceptual framework for an assessment needs to clearly articulate the cognitive demands of the task, problem-solving skills that can be used, and criteria to judge performance.

Assessment Design

Well-designed test specifications are more important for performance assessments than for multiple-choice tests because there are fewer performance tasks, and each is typically designed to measure, in part, something that is unique (Haertel & Linn, 1996). Test specifications delineate the content, cognitive processes, and statistical characteristics of the assessment tasks. The test specifications, purpose of the assessment, population of potential examinees, and the intended score interpretations guide the design of the assessment tasks.

The deeper the understanding is of how students acquire and structure knowledge and cognitive skills and of how they perform cognitive tasks, the better we are at assessing students' cognitive thinking and reasoning skills and obtaining information that will lead to improved learning. Cognitive theories of learning are needed to design assessments that can be used in meaningful ways to guide instruction and monitor student learning. A systematic approach to designing assessments that reflects theories of cognition and learning is evidence-centered design (Mislevy, Steinberg, & Almond, 2003), in which evidence observed in student performances on complex problem-solving tasks that have clearly

articulated cognitive demands is used to make inferences about student achievement and learning.

Performance assessment design should consider the degree of structure for the problem posed and the response expected. Baxter and Glaser (1998) characterized performance assessments along two continuums with respect to their task demands. One continuum reflects the task demand for cognitive processes ranging from open to constrained, and the other continuum represents the task demand for content knowledge ranging from rich to lean. If a performance task allows for opportunities for students to develop their own strategies and procedures, it is considered process open. If it requires substantial content knowledge for successful performance, it is considered content rich. By crossing these two continuums, four quadrants are formed so that tasks can be designed to fit one or more of these quadrants, which can result in clearly articulated cognitive and content targets for task design and for evaluation of tasks in terms of their alignment with these targets (Baxter & Glazer, 1998). The two continuums could allow for more than four quadrants to allow for examination of students' progression in understanding within a content domain.

Templates for task design. A task template can guide the design of tasks that assess the same cognitive skills, and a scoring rubric can then be designed for the family of tasks generated by a particular template. The use of templates can ensure that the intended cognitive skills are assessed and can thus improve the generalizability of score inferences. As Baker (2007) has suggested, cognitive task demands can be represented by families of tasks or templates such as reasoning, problem solving, and knowledge representation tasks. For example, an explanation task template requires students to read one or more texts, requiring some prior knowledge of the subject for students to understand the text and to evaluate and explain important issues and concepts addressed in the text (Niemi, Baker, & Sylvester, 2007).

Computer-based simulation task design. Computer-based simulation tasks allow for the assessment of complex reasoning and problem-solving skills that cannot be measured using more traditional assessment formats. They can assess students' skills in formulating, testing, and evaluating hypotheses; selecting appropriate solution strategies; and if necessary adapting strategies on the basis of the degree of success to solution. Students' strategies, as well as their products, can be captured, which can be valuable in monitoring the progression of student learning and guiding instruction (Bennett, Persky, Weiss, & Jenkins, 2007). Automated scoring procedures for evaluating student performances on computer-based simulation tasks allow for timely feedback and address the cost and demands of human scoring.

As is the case with all assessments, computer-based tasks have the potential to measure factors that are irrelevant to the target assessment construct, and therefore the validity of the score interpretations may be jeopardized. Examinee familiarity and practice with the computer interface is important. It is also essential to ensure that that the range of cognitive skills and knowledge assessed by the computer-based tasks are not narrowed to those that are easily assessed using computer technology and that the automated scoring procedures reflect important features of student achievement so that the generated scores allow for accurate interpretations (Bennett & Gitomer, 2006). The use of task templates can help ensure that the breadth of the domain is assessed and the scoring rubrics embody the important cognitive demands.

Use of learning progressions in assessment design. Assessment design can be guided by learning progressions that indicate what it means to acquire more expert understanding within subject domains. Learning progressions are based on models of cognition and learning, but for many subject domains cognitive models of how competency develops have not been fully developed (Wilson & Bertenthal, 2005). The specification of learning progressions may therefore be supplemented by what expert teachers know about student learning. Learning progressions based on cognitive models of student achievement can inform the design of assessments that will elicit evidence to support inferences about student performance at different points along the learning progression (Wilson & Bertenthal, 2005).

Carefully crafted performance assessments are capable of capturing student understanding along these learning progressions.

Design of Scoring Rubrics

Similar to the design of performance tasks, the design of scoring rubrics is an iterative process and involves coordination across grades (Lane & Stone, 2006). Scoring rubrics may not be unique to a specific task or generic to the entire construct domain, but they may be reflective of a family of tasks or a particular task template, or the assessment may have a generic scoring rubric as well as task-specific rubrics that are aligned with and reflect important features of the generic rubric (Lane & Stone, 2006). In designing scoring rubrics, many factors need to be considered, including the criteria for judging the quality of performances, the choice of a scoring procedure such as an analytic or holistic method, ways for developing criteria, and whether trained raters or computer-automated scoring procedures will be used (Clauser, 2000).

Specifying scoring criteria. When specifying the criteria for judging student performances, several factors need to be considered, including the cognitive demands of the performance tasks, the degree of structure or openness intended in the response, the examinee population, the purpose of the assessment, and the intended score interpretations (Lane & Stone, 2006). The knowledge and skills reflected at each score level should differ distinctly from those at other score levels; thus, the number of score levels used depends on the extent to which the criteria across the score levels can reliably differentiate student performances. Learning progressions can also be reflected in the criteria so as to identify what skills and knowledge students have acquired and to monitor their progression.

Scoring procedures. Typically, either a holistic or an analytic approach is adopted for scoring. For holistic scoring, raters make a single, holistic judgment regarding the quality of the response and assign one score on the basis of the scoring criteria at each score level and benchmark papers that are anchored at each score level. Analytic scoring requires that the rater evaluate the response according to several dimensions, a score is provided for each of the dimensions, and the scores can then be summed to arrive at a total score. For analytic scoring rubrics, criteria are identified at each score level for each of the dimensions of interest, such as mechanics, voice and focus, and organization in a writing assessment.

Human scoring. Student responses to performance assessments are evaluated by human scorers or automated scoring procedures that have been informed by human scoring. An overview of the training procedures and methods for human scorers, the design of rating sessions that may involve raters spending several days together, and the procedures for online rating of student work is provided in Lane and Stone (2006). A major consideration in human scoring is rater variability or inconsistency. Raters may differ in the extent to which they implement the scoring rubric; the way in which they interpret the scoring criteria; the extent to which they are severe or lenient in scoring student performances; their understanding and use of scoring categories; and their consistency in rating across examinees, scoring criteria, and tasks (Eckes, 2008). As a result of rater inconsistencies, the construct representation of the assessment can be jeopardized by raters' interpretation and implementation of the scoring rubric and by features specific to the training session. Carefully designed scoring rubrics and training procedures can help alleviate errors in human scoring (Lane, 2010).

Automated scoring systems. Computer-based performance assessments, such as computer-delivered writing assessments and computer-based simulation tasks, are typically scored by automated scoring systems that have been informed by human scoring. Automated scoring procedures have several attractive features: (a) They are consistent in their application of the scoring rubric, (b) they explicitly allow for the test designer to control what features are attended to in scoring student responses, (c) they allow for the collection and recording of multiple features of student performance, and (d) they allow for scores to be generated and reported in a timely manner (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Williamson, Behar, & Mislevy, 2006).

Validation studies for automated scoring systems can provide evidence for appropriate score

interpretations and uses. The three categories of validation approaches for automated scoring systems are (a) approaches focusing on the consistency among scores given by different scorers (human and computer), (b) approaches focusing on the relationship between test scores and external measures of the construct being assessed, and (c) approaches focusing on the scoring process (Yang, Buchendahl, Juszkiewicz, & Bhola, 2002). Studies have focused on the first category, on the relationship between human- and computer-generated scores, typically indicating that the relationship between human scores and computer-generated scores is very similar to the relationship between the scores produced by two humans. Validation studies that focus on the latter two categories are scarce. Moreover, both construct-irrelevant variance and construct underrepresentation may affect the validity of scores generated by automated scoring systems in that the systems may be influenced by irrelevant features of students' responses and assign a higher or lower score than deserved or the systems may not fully represent the construct being assessed, which can affect score interpretations (Powers et al., 2002).

VALIDITY OF PERFORMANCE ASSESSMENTS

In the evaluation of performance assessments, evidence to support the validity of the score inferences is at the forefront. *Validity* pertains to the meaningfulness, appropriateness, and usefulness of test scores (AERA et al., 1999; Kane, 2006; Messick, 1989). Assessment validation requires the specification of the purposes and uses of the assessment, the design of an assessment that fits the intended purposes, and the collection of evidence to support the proposed uses of the assessment and the intended score inferences.

Two sources of potential threat to the validity of score inferences are construct underrepresentation and construct-irrelevant variance (Messick, 1989). Construct underrepresentation occurs when an assessment does not fully capture the targeted construct, resulting in score inferences that may not be generalizable to the larger domain of interest. Construct-irrelevant variance occurs when one or more irrelevant constructs is being assessed along with the intended construct. Sources of construct-irrelevant variance for performance assessments may include task wording, task context, response mode, and raters' attention to irrelevant features of responses. As an example, if a science performance assessment requires a high level of reading ability and students who have very similar science proficiency perform differently because of differences in their reading ability, the assessment is in part measuring a construct that is not the assessment target—reading proficiency. Students' writing ability could be a source of construct-irrelevant variance for tasks that require students to explain their reasoning on mathematics and science assessments. These examples illustrate that construct-irrelevant variance may hinder the performance of subgroups of examinees such as English language learners and is therefore inherently related to the fairness of assessments for subgroups of examinees (see the Fairness of Assessments section). Construct-irrelevant variance may also occur when raters score student responses to performance tasks according to features that do not reflect the scoring criteria and are irrelevant to the construct being assessed (Messick, 1994). This variance can be addressed by clearly articulated scoring rubrics and effective training of the raters.

Validity criteria that have been suggested for examining the quality of performance assessments include content representation, cognitive complexity, meaningfulness, generalizability, fairness, and consequences (Linn, Baker, & Dunbar, 1991; Messick, 1994). These criteria are closely aligned to the sources of validity evidence specified in the *Standards for Educational and Psychological Measurement* (AERA et al., 1999).

Evaluating Content Representation

The alignment between the content of the assessment and the construct it is intended to measure provides validity evidence of score inferences (AERA et al., 1999). *Test content* refers to the skills, knowledge, and processes that are intended to be assessed by tasks as well as to the task formats and scoring procedures. Although performance tasks may be assessing student understanding of concepts at a deeper level, the content of the domain may not be

well represented by a relatively small subset of performance tasks. Thus, it is important to make sure that the ability to generalize from a student's score on a performance assessment to the broader domain of interest is not limited by having too few tasks. One method to address lack of generalizability is to include other item formats that can appropriately assess certain skills and to use performance tasks to assess complex thinking skills that cannot be assessed by the other item formats.

The coherency among the assessment tasks, scoring rubrics and procedures, and target domain is another aspect of validity evidence for score interpretations. It is important to ensure that the cognitive skills and content of the target domain are systematically represented in the tasks and scoring procedures. The method used to transform performance into a score also provides evidence for the validity of the score interpretation.

Evaluating Cognitive Complexity

An attractive aspect of performance assessments is that they can be designed to assess complex thinking and reasoning skills that cannot be easily measured by other assessment formats, but as Linn et al. (1991) have cautioned, one should not assume that a performance assessment measures complex thinking skills. Evidence is needed to examine the extent to which tasks and scoring rubrics capture the intended cognitive skills. The use of cognitive theories of student learning in the design of performance assessments will enhance the validity of score interpretations. Several methods have been used to examine whether tasks are assessing the intended cognitive skills and processes (Messick, 1989), including protocol analysis, analysis of reasons, and analysis of errors. Students are asked to think aloud as they solve a problem or describe retrospectively how they solved the problem in protocol analysis. In the analysis-of-reasons method, students are asked to provide rationales for their responses to the tasks. The analysis-of-errors method requires an analysis of the procedures, concepts, or representations of the problems so as to make inferences about students' misconceptions in their thinking. For example, in the design of a science performance assessment, Shavelson and Ruiz-Primo (1998) used protocol analysis to compare expert and novice reasoning on the tasks. The protocol analysis results confirmed several of their hypotheses regarding the different reasoning skills that the tasks were intended to elicit, illuminated the complexity of experts' reasoning compared with that of novices, and informed the design of tasks and interpretation of scores.

Evaluating Meaningfulness and Transparency

An important validity criterion for performance assessments is their meaningfulness (Linn et al., 1991), which refers to the extent to which students, teachers, and other interested stakeholders find value in the tasks. A related criterion is transparency—students and teachers know what is being assessed, by what methods, the criteria used to evaluate performance, and what is successful performance (Frederiksen & Collins, 1989). It is important to ensure that all students are familiar with the format of tasks and the scoring criteria. As part of instruction, teachers can use performance tasks with their students and engage them in discussions about what the tasks are assessing and the nature of the criteria used for evaluating student work. Students can use scoring rubrics to evaluate their own work and that of their peers.

Evaluating the Generalizability of Score Inferences

A potential threat to the validity of score interpretations is the extent to which the scores from the performance assessments can be generalized to the broader construct domain (Linn et al., 1991). Generalizability theory provides both a conceptual and a statistical framework to examine the extent to which scores derived from an assessment can be generalized to the domain of interest (Brennan, 2001). Furthermore, generalizability theory can inform the design of an assessment system so as to ensure the validity of score inferences. For example, it can provide information on the number of items, raters, and occasions that are needed to maximize generalizability of scores for absolute or relative decisions, or both. It is particularly relevant in evaluating performance assessments because it examines multiple

sources of error that can limit the generalizability of scores, such as error resulting from tasks, raters, and occasions. Error resulting from tasks occurs because only a few tasks are typically included in a performance assessment. Students' individual reactions to specific items tend to average out on multiple-choice tests because of the relatively large number of items, but such individual reactions to items have a greater impact on scores on performance assessments consisting of fewer items (Haertel & Linn, 1996). It is important to consider the sampling of tasks, and by increasing the number of tasks on an assessment, the validity and generalizability of assessment results can be enhanced. The use of multiple item formats, including performance tasks, can improve the generalizability of the score inferences.

Error resulting from raters can also affect the generalizability of scores in that raters may differ in their evaluation of the quality of students' responses to a particular performance task and across performance tasks. Raters can differ in their stringency, resulting in rater mean differences, and they can differ in their judgments about whether one student's response is better than another student's response, resulting in an interaction between the student and rater facets (Lane, Liu, Ankenmann, & Stone, 1996; Shavelson, Baxter, & Gao, 1993). Occasion is an important hidden source of error because performance assessments are typically only given on one occasion, and occasion is generally not considered in generalizability studies (Cronbach, Linn, Brennan, & Haertel, 1997).

Generalizability studies have shown that error resulting from raters for hands-on science performance tasks (e.g., Shavelson et al., 1993) and for mathematics performance tasks (Lane et al., 1996) tends to be smaller than that for writing assessments. To help obtain consistency among raters, attention is needed in the design of scoring rubrics, selection and training of raters, and evaluation of rater performance before and throughout operational scoring of student responses (Lane & Stone, 2006). Researchers have shown that task-sampling variability in students' scores is a greater source of measurement error in science, mathematics, and writing performance assessments than rater-sampling variability; therefore, increasing the number of tasks has a greater effect on the generalizability of the scores than increasing the number of raters (Lane et al., 1996; Shavelson et al., 1993).

Lane et al. (1996) showed that task-sampling variability was the major source of measurement error on a mathematics performance assessment that required students to show their solution processes and explain their reasoning. The results indicated that error resulting from raters was negligible, whereas error resulting from tasks was more substantial, indicating differential student performance across tasks. Generalizability studies (Lane et al., 1996; Shavelson et al., 1993, 1999) indicated that between 42% and 62% of the total score variation was accounted for by the Person × Task variance component, indicating that people were responding differently across tasks because of task specificity. The variances resulting from the rater effect, the Person × Rater interaction, and the Rater × Task interaction were negligible. When the number of tasks was equal to 9, the generalizability coefficients for student scores ranged from .71 to .84. The coefficients for absolute decisions for school-level scores ranged from .80 to .97 when the number of tasks was equal to 36 using a matrix sampling design. These results provided evidence that the assessment allowed for accurate generalizability of school-level scores.

Shavelson et al. (1993) provided evidence that the large task-sampling variability in science performance assessments was the result of variability not only in the Person × Task interaction, but also in the Person × Task × Occasion interaction. The Person × Task variance component accounted for 32% of the total variability, whereas the Person × Task × Occasion variance component accounted for 59%. The latter result suggests that students performed differently on each task from occasion to occasion. Shavelson, Ruiz-Primo, and Wiley (1999) provided additional support for the large effects resulting from occasion in that the Person × Task variance component accounted for 26% of the total variability and the Person × Task × Occasion variance component accounted for 31% of the total variability, indicating a tendency for students to change their approach to each task from occasion to occasion. The variance component for the Person × Occasion

effect was close to zero. Shavelson et al.'s results indicate that although students approached the tasks differently on different testing occasions, once the data were aggregated over the tasks, their aggregated performance did not vary from one occasion to another. It is important to note that the person variance component accounted for only 4% of the total variability, indicating that the variability of the overall scores for people was relatively small. As Shavelson et al. (1999) indicated, the sample of students was homogeneous and generally scored very high on the tasks.

The results from generalizability studies (Lane et al., 1996; Shavelson et al., 1993, 1999) have indicated that scoring rubrics and the procedures used to train raters can be designed so as to minimize rater error, and increasing the number of performance tasks will increase the generalizability of the scores. Likewise, including other item formats on performance assessments will facilitate the generalizability of scores to the broader domain. The reader is referred to Chapter 3 of this volume.

Fairness of Assessments

The evaluation of an assessment's fairness is inherently related to all sources of validity evidence. Bias can be conceptualized "as differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (Cole & Moss, 1989, p. 205). A fair assessment requires evidence to support the meaningfulness, appropriateness, and usefulness of the test score inferences for all relevant subgroups of examinees. Validity evidence for assessments that are intended for students from various cultural, ethnic, and linguistic backgrounds needs to be collected continuously and systematically as the assessment is being developed, administered, and refined. When sample sizes permit, analyses performed on the entire sample of examinees (e.g., factor analyses, item analyses, predictive validity studies) should also be performed on subgroups of examinees to evaluate the comparability of the construct being assessed across groups. Differential item functioning should also be examined to ensure that examinees of equal ability, regardless of their subgroup affiliations, are performing similarly on items.

The intended population and subpopulations of examinees should be considered in the design of assessments. As an example, the linguistic demands of items can be simplified to help ensure that English language learners are able to access the task as well as other students. Abedi and Lord (2001) have demonstrated that by simplifying the linguistic demands of items, the gap between English language learners and other students can be narrowed. The contexts used in performance tasks can be evaluated to ensure that they are familiar to various subgroups and will not negatively affect the task performance of one or more subgroups. The amount of writing required on mathematics, reading, and science assessments, for example, can be examined to help ensure that writing ability will not unduly influence students' ability to demonstrate what they know and can do on these assessments. Scoring rubrics can be designed to ensure that the relevant math, reading, or science skills are the focus, not students' writing ability. The use of other response formats, such as graphic organizers, on reading assessments may alleviate the concerns about writing ability confounding student performance on reading assessments (O'Reilly & Sheehan, 2009). Chapter 17, this volume, and Volume 3, Chapter 17, this handbook, address fairness issues in assessment more fully.

Consequential Evidence

The evaluation of both intended and unintended consequences of any assessment is fundamental to the validation of score interpretation and use (Messick, 1989, 1994). Because a major rationale for performance assessments is to improve instruction and student learning, obtaining evidence of such positive consequences and any potentially negative consequences is even more compelling (Linn, 1993; Messick, 1994). Moreover, adverse consequences bearing on issues of fairness are particularly relevant because one cannot assume that a contextualized performance task is equally appropriate for all students because "contextual features that engage and motivate one student and facilitate his or her effective task performances may alienate and confuse another student and bias or distort task performance" (Messick, 1994, p. 19). This concern can be dealt with by using a well-specified

design process in which fairness issues are addressed, including expert analyses of tasks and rubrics and analyses of student thinking as they solve performance tasks, with special attention to examining potential subgroup differences and features of tasks that may contribute to these differences.

Performance assessments that measure complex thinking skills have been shown to have a positive impact on instruction and student learning (Lane, Parke, & Stone, 2002; Stecher, Barron, Chun & Ross, 2000; Stone & Lane, 2003). In a study examining the consequences of Washington's state assessment, approximately two thirds of fourth- and seventh-grade teachers reported that the state standards and the state assessment short-answer and extended-response items were influential in improving instruction and student learning (Stecher et al., 2000). The examination of the relationship between changes in instructional practice and improved student performance on the performance assessments is an important aspect of consequential evidence for performance assessments. The relationship between changes in instructional practice and improved performance on the Maryland State Performance Assessment Program (MSPAP), which consisted entirely of performance tasks that were integrated across content domains, was examined by Lane et al. (2002; Stone & Lane, 2003). The results indicated that teacher-reported reform-oriented instructional features accounted for differences in school performance on the MSPAP in reading, writing, mathematics, and science. They also indicated that schools in which teachers reported that their instruction over the years reflected more reform-oriented problem types and learning outcomes similar to those assessed by the MSPAP had higher levels of school performance on the MSPAP than schools in which teachers reported that their instruction reflected less reform-oriented problem types and learning outcomes. The results also indicated that schools in which teachers reported that their instruction over the years reflected more reform-oriented problem types and learning outcomes accounted for differences in the rate of change in MSPAP school performance in reading and writing. These results suggest increased reported use of reform-oriented performance tasks in writing and reading, and a focus on the reading and writing learning outcomes in instruction was associated with greater rates of change in MSPAP school performance over a 5-year period.

When using test scores to make inferences regarding the quality of education, contextual information is needed to inform the score inferences and actions (Haertel, 1999). In this regard, Stone and Lane (1993) showed that a school contextual variable, percentage of students eligible for free or reduced-cost lunch, which is typically used as a proxy for socioeconomic status, was significantly related to school-level performance on the MSPAP in mathematics, reading, writing, science, and social studies. Schools with a higher percentage of students eligible for free or reduced-cost lunch tended to perform poorer on the MSPAP. More important, no significant relationship was found between percentage of students eligible for free or reduced-cost lunch and growth on the MSPAP at the school level in four of the five subject areas—mathematics, writing, science, and social studies. This result indicated that school-level growth on the science, math, writing, and social studies performance assessment was not related to the percentage of students who were eligible for free or reduced-cost lunches within the school.

CONCLUDING THOUGHTS

The educational benefit of using performance assessments has been demonstrated by many researchers and educators. When students have had the opportunity to work on meaningful, real-world tasks in instruction, they have demonstrated improved performance on assessments that reflect these kinds of tasks. The Common Core State Standards and Race to the Top initiatives call for the alignment of curriculum, instruction, and assessment as well as the need to instruct and assess students on complex thinking skills that will prepare them for college and careers. Ample evidence has supported the use of performance assessments in both instruction and assessment to improve achievement and learning for all students.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219–234. doi:10.1207/S15324818AME1403_2

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Baker, E. L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment, 12,* 179–194.

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48,* 1210–1218. doi:10.1037/0003-066X.48.12.1210

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17,* 37–45. doi:10.1111/j.1745-3992.1998.tb00627.x

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York, NY: Springer. doi:10.1007/978-1-4020-9964-9_3

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project* (NCES 2007–466). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24,* 310–324. doi:10.1177/01466210022031778

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–220). New York, NY: Macmillan.

Council of Chief State School Officers and National Governors Association. (2010). *Common Core Standards for mathematics*. Retrieved from http://www.corestandards.org

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373–399. doi:10.1177/0013164497057003001

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of school and students at work*. New York, NY: Teachers College Press.

Darling-Hammond, L., & Pecheone, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning.* Retrieved from http://www.k12center.org/rsc/pdf/Darling-HammondPechoneSystemModel.pdf

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25,* 155–185. doi:10.1177/0265532207086780

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*, 27–32.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan, 80*, 662–667.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (NCES 96–802, pp. 59–78). Washington, DC: U.S. Department of Education.

Kane, M. T. (2006). Validation. In B. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Lane, S. (2010). *Performance assessment: The state of the art*. Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from http://scale.stanford.edu/system/files/performance-assessment-state-art.pdf

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*, 71–92. doi:10.1111/j.1745-3984.1996.tb00480.x

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning. *Educational Assessment, 8*, 279–315. doi:10.1207/S15326977EA0804_1

Lane, S., & Stone, C. A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–432). Westport, CT: Praeger.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1–16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15–21.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13–23.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62. doi:10.1207/S15366359MEA0101_02

Niemi, D., Baker, E. L., & Sylvester, R. M. (2007). Scaling up, scaling down: Seven years of performance assessment development in the nation's second largest school district. *Educational Assessment, 12*, 195–214.

No Child Left Behind Act of 2001, Pub. L. 107–110, 115 Stat. 1425 (2002).

O'Reilly, T., & Sheehan, K. M. (2009). *Cognitively based assessment of, for and as learning: A framework for assessing reading competency* (ETS RR-09-26). Princeton, NJ: Educational Testing Service.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28,* 5–13. doi:10.1111/j.1745-3992.2009.00149.x

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing, 26*, 407–425. doi:10.1092/UP3H-M3TE-Q290-QJ2T

Resnick, L. B., & Resnick, D. (1992). Assessing the thinking curriculum. In B. B. Gifford & M. C. O'Conner (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37–75). Boston, MA: Kluwer Academic.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215–232. doi:10.1111/j.1745-3984.1993.tb00424.x

Shavelson, R. J., & Ruiz-Primo, M. A. (1998). *On the assessment of science achievement conceptual underpinnings for the design of performance assessments* (CSE TR 481). Los Angeles, CA: CRESST.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability. *Journal of Educational Measurement, 36,* 61–71. doi:10.1111/j.1745-3984.1999.tb00546.x

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000, August). *The effects of the Washington state education reform in schools and classrooms* (CSE Tech. Rep. No. 525). Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16,* 1–26. doi:10.1207/S15324818AME1601_1

U.S. Department of Education. (2005). *The nation's report card.* Washington, DC: Author. Retrieved from http://nationsreportcard.gov/science_2005/s0116.asp

U.S. Department of Education. (2009). *Race to the Top Program executive summary*. Retrieved from http://www.ed.gov/programs/racetothetop/resources.html

Williamson, D. M., Behar, I. I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 1–13). Mahwah, NJ: Erlbaum.

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2005). *Systems for state science assessment*. Washington, DC: National Academies Press.

Yang, Y., Buchendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*, 391–412.

CHAPTER 21

LANGUAGE TESTING: HISTORY, VALIDITY, POLICY

Tim McNamara

Many research and real-world contexts demand a determination of how well a person can understand and speak a language. The research field of language testing has emerged in response to this need. It is a cross-disciplinary field, drawing principally on applied linguistics and psychometrics. Perhaps more so than in other specific fields of measurement, given the complexity of language, domain expertise (in theories of language, language learning, and language use, particularly of second and additional languages) is as important as measurement expertise in the effective development of language tests.

THE CONSTRUCT IN LANGUAGE TESTS: LANGUAGE, LANGUAGE USE, AND LANGUAGE LEARNING

The domain of language tests, and the resulting test construct, reflect theories of language in linguistics, of language use in sociolinguistics and linguistics pragmatics, of language processing in psycholinguistics, and of language learning in the field of language acquisition research. Developments in each of these fields have had successive impacts on the design and validation of language tests. The traces of each of these developments can still be felt in language testing practice, which reflects a range of conservative and more up-to-date influences.

The advent of modern linguistics in the work of the Swiss linguist Ferdinand de Saussure (1983) was a watershed in the development of language teaching and of language tests. Before this, the discipline of linguistics, then known as *philology,* was dominated by studies of the history of language, language change, and comparison of languages within and across language families. Language teaching was modeled on the teaching of Greek and Latin, and language tests featured translation and composition, with little attention to the spoken language. This *diachronic* approach, that is, focusing on the evolution of language systems over time, was replaced in de Saussure's work by an emphasis on a *synchronic* description of language, that is, as a system at a given point in time, without concern for how it arrived at that state. This emphasis was reflected in a focus on the description of what people actually said and did in language rather than on the prescription of what people ought to say or do in language to be "correct." Language in Saussurean linguistics was understood as a set of systems of structural contrasts at the level of phonology (pronunciation), morphology (word formation), syntax (the rules of grammatical combination), and semantics (lexical or word meaning). De Saussure gave the name *langue* to the system of language shared among all the speakers within a speech community and contrasted it with *parole,* the actual use of language in context, which he saw as too complex for systematic study.

One very important effect of the impact of Saussurean linguistics was a focus on the description of spoken language, which was important particularly for linguistic anthropologists in the United States working on the languages of nonliterate cultures, including Native American languages, which could be accurately described and analyzed for the first time. Linguists trained in the structuralist

DOI: 10.1037/14047-021
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

tradition played a prominent role in the teaching of the spoken languages of the theaters of war to U.S. Army personnel during World War II (Spolsky, 1996). The establishment of the English Language Institute at the University of Michigan in 1941 by the structuralist linguist Charles Fries saw a similar emphasis on the spoken language and structuralist approaches in the teaching of English as a foreign language to students from non-English-speaking countries wanting to study, usually for a higher degree, in the United States. The positive reputation of the so-called Army method meant that when, after the Sputnik crisis of 1957, the U.S. government authorized massive spending on improving the teaching of foreign languages, the method used, audiolingualism, which focuses on the mastery of the systematic features of language underlying spoken communication, was the preferred teaching method, delivered in the newly available language laboratories. This method was supported by behaviorist theories of learning; spoken language patterns were imitated and memorized.

Language tests in this early period of modern, scientific language testing, termed the *psychometric–structuralist period* by Spolsky (1978), were marked by separate, objective subtests of mastery of the individual systems of language—grammar (syntax and morphology), vocabulary, and phonology (particularly sound contrasts). Fries's colleague and eventual successor at the University of Michigan's English Language Institute, Robert Lado, developed a program in language testing there. He applied the insights of structural linguistics to the design and development of language tests, and his methods were described in his seminal work *Language Testing* (Lado, 1961). Multiple-choice tests, whose reliability had been carefully established using psychometric methods, were the means of choice. Tests of writing and speaking as whole skills were viewed with suspicion because reliable assessments of performance in these skills, which needed to be judged subjectively, were recognized to be difficult to achieve.

At the same time that Lado's (1961) methods were becoming known and adopted, the educational psychologist John Carroll gave an important paper (Carroll, 1961/1972) stressing the need to complement this atomistic testing of the elements of language knowledge with the integrated testing of skills in performance. Carroll's remarks foreshadowed the subsequent major developments in language testing that emerged a decade later, each of which emphasized an integrated performance. The new approach conceptualized competence in a language as a skill, such as being able to ride a bicycle or play an instrument, not simply as a field of knowledge as with academic subjects such as history or mathematics. Theories of the acquisition of language supported this change, seeing it as a natural, largely unconscious process and as involving gradually increasing coordination of underlying psycholinguistic processes in reception and production. These developments had revolutionary impacts on the nature and design of language tests, which came to reflect one of two contrasting ways of conceptualizing the notion of what was called *performance* in languages. (The distinction between *performance* and *competence,* or underlying knowledge of the systems of language, was introduced by Chomsky, 1965.)

The first way, developed in the work of John Oller (1979), saw performance in cognitive, psycholinguistic terms. He suggested that tests should reflect the conditions of performance: integration of various areas of language knowledge (morphology, lexicon, syntax, and in oral tests, phonology) within an unfolding textual context and, for speech reception and production, in real time. Oller argued that oral tests such as dictation and reading tests such as cloze (in which passages are presented with selected words deleted, the task being to supply the missing words) met this criterion. Such tests were called *integrative tests,* in contrast to the separate discrete-point tests of Lado (1961). Note that the tasks Oller recommended (dictation, cloze) did not resemble real-world tasks. Oller argued that performance on any one of these tasks would be predictive of performance on others, because what was being tapped was an integrative performance capacity, something not captured in discrete-point tests. Oller's claims were subsequently disputed in research, but the attention to the underlying cognitive skills involved in performance was original and has been returned to in recent developments. Modified forms of cloze testing remain popular to this day as general proficiency measures; a related procedure is the C-test, in

which the second half of every second word in a text is deleted and must be supplied (Klein-Braley, 1997).

The second and more influential approach to performance was more sociolinguistic in character and addressed de Saussure's (1983) *parole,* the actual use of language in context, more directly. It drew on the performance assessment tradition developed in the U.S. Army during World War II, when individuals needed to be allocated to tasks requiring particular practical skills. To assess whether an individual was competent to perform a skill—repairing equipment, for example—the task involved was simulated in the test setting and the individual was asked to perform it. The performance would be observed and judged. A very influential test of speaking emerged with the same motivation in the 1950s, when the Foreign Service Institute of the U.S. Department of State in Washington, DC, needed a means of assessing the spoken communication skills of personnel who were to be allocated to a range of duties overseas requiring different, sometimes very demanding levels of communicative skill. Unfortunately, the theory of language testing available at the time (objective discrete-point testing, which eschewed the subjective testing of spoken language) offered no useful guidance. Therefore, the experience of those who had already performed successfully in the field was drawn on to define a scale of five broad levels of increasing complexity of communicative task and increasing quality of performance, up to performances at the educated-native-speaker level. An oral proficiency interview was used to elicit a performance, the qualities of which were matched against the scale (Clark & Clifford, 1988). This interview was the first example of a criterion-referenced language test and has been hugely influential, providing a template for many tests of spoken competence ever since.

Subsequent developments provided a somewhat belated theoretical basis for this tradition of performance assessment as part of what came to be known as the communicative movement in language teaching and testing, still the predominant orthodoxy to this day. The linguistic anthropologist Dell Hymes located communication in its cultural context. He proposed a model of what it is to know a language to which he gave the name *communicative competence* (Hymes, 1972b). Hymes's model incorporated Chomsky's (1965) cognitive model of linguistic competence and extended it to an expanded notion of performance that included psycholinguistic, interactional, and sociolinguistic dimensions. In addition, he set out a methodology for analyzing culturally and socially specific events involving language behavior (Hymes, 1972a), which enabled the specification of competence to manage the communicative demands of such events. Hymes's work was reinterpreted for second language settings by Canale and Swain (1980), which in turn inspired the influential Bachman (1990) model of communicative language ability. This model attempted to define the cognitive demands of communication by setting out the components of individual ability needed for successful performance of communicative tasks, which then form the basis for a template for designing appropriate test tasks. More and more, language tests attempted to define the target language use context and to replicate it in the test. This method has been prominent in testing the English of international students wanting to study in universities conducted in the medium of English. The best known examples are the Internet-Based Test of English as a Foreign Language, introduced by the Educational Testing Service in 2005 to replace its previous largely discrete-point test, little changed since the 1960s (Chapelle, Enright, & Jamieson, 2008), and the British–Australian International English Language Testing System, introduced more than a decade earlier (Davies, 2008a). Other specific-purpose performance tests have targeted work settings. For example, the Occupational English Test (McNamara, 1996) was designed to assess the ability of nonnative English-speaking health professionals to manage the communicative demands of the clinical workplace, which it attempts to replicate in its tasks. General-purpose language tests in the communicative tradition have also focused on replication of common communicative activities involved in conducting routine transactions, for example, while traveling or as required in intercultural social settings.

Another theoretical source for communicative language teaching and testing has been the naturalistic theories of language acquisition that appeared

in the 1960s and 1970s. These theories resulted from the critique of behaviorism that emerged at that time and the discovery of consistent sequences of acquisition, first in the mother-tongue language acquisition of children and then in second language learners, including adults. The discovery of these sequences led researchers and then teachers to realize that language learning was not a question of consciously accumulating elements of knowledge and skill, as in the structuralist approach. Instead, language learning was now understood to be a far more organic and instinctive process, which suggested that integrated performance on whole tasks should be the focus of assessment in the communicative tradition.

VALIDATION OF LANGUAGE TESTS

The tests of the psychometric–structuralist period operated with notions of test validity that emphasized domain sampling, reliability, and criterion-related validity. In contrast, discussions of the validity of communicative language tests soon focused on the more comprehensive theories of validity developed within educational assessment, particularly in the work of Samuel Messick (1989; Bachman, 1990; McNamara, 2006), more recently as interpreted by Michael Kane (2001; Bachman, 2005; see also Chapter 4, this volume). Before this, communicative tests, particularly in the British tradition, had focused validation efforts on evidence of how test content represented tasks in the real-world domain of interest (e.g., Weir, 1983, on the validation of a test of English for academic purposes). This approach was open to critique after Messick, who famously likened claims of the validity of tests on the basis of their content to the barker at the circus making claims about circus freaks—evidence of the reality of the claims was lacking. In the new approach, attention was paid to steering between the Scylla and Charybdis of construct underrepresentation and construct-irrelevant variance.

Construct Underrepresentation in Language Tests

Construct underrepresentation is a severe threat in language tests, because it is difficult to replicate and sample real-world conditions of language performance in the artificial context of the test setting. The design of test formats in the interests of manageability or reliability often comes at the expense of validity, which is particularly important in the testing of speaking.

Recent approaches to defining the construct of speaking embrace the view that speaking is not a solo performance. Instead, spoken interaction is seen as a joint construction, a carefully coordinated effort in which the performance of one person in the interaction is dependent on the performance of the other—much as in dancing, in which with a good partner one may seem an accomplished dancer, but with a clumsy partner one's efforts may seem awkward. In terms of task design, this new orientation would suggest that assessments of speaking should involve two or more participants. On the grounds of cost, however, monologic tests of speaking, using digital means of delivery of stimulus and capture of performance, may be preferred. This, of course, raises questions of the comparability of the two formats (interactive and monologic); the implications for score meaning have been studied by O'Loughlin (2001). Studies of interaction in the most frequently used format for assessing speaking ability, the Oral Proficiency Interview, have also raised issues of construct underrepresentation. Discourse studies (Young & He, 1998) have shown that although the Oral Proficiency Interview is intended to replicate casual conversation, its structure and the roles the examiner and the candidate play do not resemble, in very significant ways, those of conversational participants. For example, it is unusual in such interviews for the candidate to take the initiative by asking questions or probing meaning or intent. This imbalance in the roles of interviewer and interviewee may have practical consequences in the inferences drawn about candidates. For example, McNamara (1996) found that those hiring foreign-trained nurses to work in clinics in which the language of the workplace is English would be misled on the basis of an interview about the nurse's capacity to manage interaction with patients in actual clinical settings. There, the nurse typically has to take a strong role of initiating and persuading, often with patients who are relatively uncooperative communication partners.

To make the assessment setting in tests of general spoken proficiency less like a formal interview, tests are turning to formats in which pairs or small groups of students interact with one another without the assessor's direct involvement (Taylor & Wigglesworth, 2009). These paired or group oral exams also have the advantage of reducing assessment costs because more than one candidate can be assessed at the same time, which may make a further assessment given by a second assessor affordable. It may arguably also represent the construct of communication more adequately. As far as English as a language for international communication is concerned—for example, in international travel, business, and other work settings—it is clear that most use of English these days involves English as a lingua franca communication, that is, communication in which the conversational partner is more likely to be another nonnative English speaker than a native English speaker. Although new, uncontrolled variables are introduced—the personalities of the interlocutors and their level of proficiency relative to the candidate being assessed—it could be argued that this is in fact true to the construct of English as a lingua franca communication, in which variability among interlocutors in terms of their personality and proficiency is part of what an individual has to deal with in real-world settings. However, this concern that the test represent the conditions of authentic communication raises issues of fairness by introducing potential differences in partners' capability. The trade-offs between validity and reliability are particularly sharp in language assessment.

A further aspect of construct underrepresentation is raised by the question of the criteria used in judging performance. These criteria will typically include such things as pronunciation, accuracy of use of grammar and vocabulary, fluency, use of appropriate levels of formality, and the like, that is, aspects of performance with a strong linguistic focus, largely unchanged from the days of structuralism. However, as John Carroll (1954) pointed out in a very early paper, these criteria underrepresent what counts in interaction. He argued that the greater part of people's capacity for spoken communication in a second language is their capacity in their first language. This insight is insufficiently addressed in the theoretical models of communicative competence underlying communicative language assessments. In particular, although the model of communicative competence proposed by Hymes (1972b) does include a range of personality factors as affecting competence, this emphasis was not reflected in the subsequent interpretations and development of Hymes's work by Canale and Swain (1980) and Bachman (1990). These authors focused on purely cognitive aspects of language ability, albeit now including sociolinguistic and strategic dimensions relevant to communication. The narrowing of the construct in this way may have practical consequences in assessment, especially in workplace settings. Here the personality of the person, and his or her level of professional competence, may be factors as important in the success of the person's communicative efforts as his or her language proficiency, narrowly conceived (Cameron & Williams, 1997; Kim & Elder, 2009). Studies have also shown that when attempts are made to direct judges' attention to more construct-relevant criteria, such as overall communicative effectiveness, criteria such as grammatical accuracy may be covertly imported by raters into the assessment as a proxy for this overall judgment, no doubt on account of the strong residual influence of structuralism in language teaching (McNamara, 1990).

Another way in which constructs may be underrepresented is apparent in the testing of listening. Most listening goes on in face-to-face interaction. However, evidence for the success or otherwise of listening in this context is hard to achieve from observation alone, unlike the success of speaking. Although failure of mutual understanding is sometimes apparent in the progress of the interaction, people tend to let things pass that they have not understood, unless these things become an explicit issue in the interaction, in which case the failure of mutual understanding does become obvious. As a result of this difficulty, listening tests typically take the form of requiring candidates to listen to speech (either monologue or multiparty interactions) in which they are not a participant. Questions are then used to establish their comprehension of key and more detailed points in the spoken material. Thus, a systematic underrepresentation of the listening

construct occurs because the social, interactive part of listening is not represented. Instead, listening is conceived of purely in cognitive processing terms.

Another problem of construct representation is raised by the conventional development of separate assessment measures for each of the so-called four macroskills (reading, writing, listening, speaking), with separate reporting of performance in each of these skills. In reality, of course, many communicative tasks require an integration of these skills. For example, a student at university may read in preparation for a lecture, listen to a lecture, discuss its content in a discussion section, and then be required to write something on the topic. Thus, at least the latter three skills are dependent on performance on earlier skills, and even the first is done with the knowledge that tasks involving the other skills will follow. (There is also the larger question of the integration of these tasks into the student's life context because the student's intellectual engagement in the real-world setting may not be easily replicated in the test setting, which is likely to have an impact on performance.) As a small move in the direction of greater fidelity to the demands of the real-world setting, language tests are beginning to use integrated tasks, in which performance in writing or speaking is dependent on prior exposure to input in the form of texts to which the candidate has listened or texts that the candidate has read.

Construct-Irrelevant Variance in Language Tests

Construct-irrelevant variance constitutes a persistent threat to the validity of language tests. Given the role of subjective judgments in performance assessment, particularly of speaking and writing, variability in judgment is a major issue. The training of judges and the calibration of judgments are important in reducing this variability as much as possible. The use of multifaceted Rasch measurement in particular, and generalizability theory to a lesser extent (Bachman, Lynch, & Mason, 1995; see also Chapter 3, this volume), have featured extensively in language testing research. These statistical techniques have enabled detailed insight into the extent and sources of variability in judgment. Although generalizability theory analyses can provide statements of the proportion of variance associated with differences among judges, it does so at the aggregate level, whereas Rasch measures can provide precise estimates of the impact of individual judges on candidates' chances of being scored in a particular category of interest. When inconsistency is the problem, the judge concerned can be retrained or ultimately even excluded from participation in the assessment. If judges are consistently harsher or more lenient than other judges, the differences among them can be allowed for in the estimate of the candidate's performance, producing fairer assessments. These statistical tools have also been invaluable in research on performance assessment. They allow one to identify possible sources of construct-irrelevant variance (e.g., in studies of the impact of task, gender, criteria, mode of task delivery) and then to design studies to estimate the impact of each of these, singly or in combination, on the candidates' chances of success. Studies of the interaction among different facets of the assessment settings—for example, the way individual raters interpret particular criteria of the rating scale—can be used to identify consistent patterns that are not compatible with the test construct. This information can then be fed back to raters in an attempt to change their behavior on subsequent rating occasions, with some limited success (Knoch, 2011).

Construct-irrelevant variance in speaking tests can also be associated with participants in the assessment setting other than raters. Brown (2005) showed the impact of the interlocutor on scores. In her study, students took the same interview test twice and were scored differently by an external rater depending on the interlocutor. Careful analysis of the discourse between the participants showed that the way the interlocutor interacted with the candidate altered the impression of the quality of the candidate's performance that was formed in the rater's mind.

Research Methods in Language Test Validation Research

Although psychometrics and its associated statistical methods were and remain the basic foundation for language testing research, the past 2 decades have seen a greater use of qualitative research methods in

language test validation. Given the background of linguistics in the training of most researchers in language testing, discourse analytic methods have increasingly been used to investigate language in language tests. Discourse studies have analyzed the language involved in stimulus texts in tests of receptive abilities (listening and reading). These studies have used methods such as measures of lexical density (the proportion of content words in relation to function words in a text) and schematic analysis (the structure of ideas within a text). Investigation of the language produced by test candidates within tests of productive abilities (speaking and writing) can also involve these methods (Brown, Iwashita, & McNamara, 2005; O'Loughlin, 1995). Studies of candidate and rater cognition often use a think-aloud or stimulated recall methodology. This methodology requires participants to articulate what they are attending to as they carry out the test-taking or test-rating task. The resulting discourse can be analyzed using thematic analysis techniques to identify recurring themes in the content. Similar methods can be used to analyze the content of spoken data from other relevant informants. For example, in tests of language targeting specific language-use settings (international aviation, clinical medicine), studies of the perceptions of instances of communication among informants who are practitioners within these settings are important. This is particularly so in view of the fact that language test developers are themselves unlikely to have the relevant experience and are thus more likely to make faulty assumptions about the character of communication in such settings relevant to the design of assessments (Long, 2005). When talk in interaction is involved, as in spoken language tests or in conversation among raters (May, 2009), conversation analysis (Schegloff, Koshik, Jacoby, & Olsher, 2002) can be a useful tool.

LANGUAGE TESTS AND POLICY

The cognitive focus of much theorizing and research on language tests has tended to obscure the use of language tests in the service of policy, at both institutional and governmental levels. This focus has in turn drawn attention away from the way in which test use can determine language test constructs and language testing practice. Messick (1989) recognized the role that values would play in test constructs and also recognized the consequences of testing policy and practice as an area needing investigation and defense as part of the process of test validation. Messick's desire to recognize a social dimension within the previously purely cognitive and psychometrically oriented world of measurement triggered intense debate within educational measurement generally (Borsboom, Mellenbergh, & Van Heerden, 2004; Kane, 2001, 2006; Popham, 1997) and also within language testing (Bachman & Palmer, 2010; McNamara & Roever, 2006) about the extent to which test use should be considered part of validity, with diverse conclusions. For example, Popham (1997) recognized test use as a fundamental issue but argued that it is not part of validity. In contrast, Borsboom et al. (2004) rejected the idea of the inclusion of test use at all and advocated a return to a purely cognitive approach. Kane (2001) wavered between seeing test use as part of validation and, in a subsequent paper (Kane, 2006), excluding it. In language testing, Bachman (2005) has proposed the idea of a test use argument as part of validation. In contrast, Shohamy (2001, 2006), in a movement she has called *critical language testing,* has stressed the fundamentally policy-driven and political character of many language tests used in educational and other settings, for example, immigration and citizenship, and the appropriation of the practice of language testing by political imperatives of dubious legitimacy.

The Biblical shibboleth test (Judges 12: 4–6) is emblematic of one of the most pervasive functions and consequences of language tests throughout history—their use as sorting and gatekeeping instruments. The word *shibboleth* was a military password that protected the lives of those who knew it and led to the slaughter of enemy soldiers who did not. Such simple, one-word shibboleth tests can be found in every age and every culture; recent examples have been noted in Sri Lanka, India, Botswana, Nigeria, and Lebanon (McNamara, 2005). Although the consequences of failure on modern, psychometrically sound tests may be less immediate and deadly (except, perhaps, for refugees, who are often subject

to a form of language test as part of establishing their identity and the legitimacy of their claims to protection; Eades, 2009), a lot is at stake in performance on such tests. Modern language tests form a precondition for promotion, employment, immigration, citizenship, or asylum. Although modern language tests are frequently administered not by sword-wielding guards but in an orderly, regulated, and lawful fashion in well-lit, carpeted test centers, they are just as much the result of political processes and value decisions as was the shibboleth test. Policies using language tests to determine who should be allowed to immigrate and who should be allowed to become a citizen or to practice their profession in another country are accompanied by debates about the deeper social and political values underlying these policies. For example, citizenship legislation in many countries requires varying levels of competence in languages held to be emblematic of national identity, and the introduction of such legislation has been accompanied by vigorous political debate (Extra, Spotti, & Van Avermaet, 2009; Hogan-Brun, Mar-Molinero, & Stevenson, 2009).

One particularly influential development is the a priori determination of language test constructs by those not involved in test construction and validation, that is, by policymakers. In an environment of increased managerialism in education, policymakers have realized that they can control the system of language education by specifying in advance the terms in which achievement will be reported. For language tests, following the communicative movement, this specification will be in terms of evidence of increasing practical communicative skill. A scale or framework consisting of an ordered, numbered series of statements of criterion-level performance provides a metric for measuring the outcomes of language education systems. The convenience and transparency of such a method has proved irresistible to educational managers and policymakers more generally (Brindley, 2008). Although a variety of scales and frameworks have been developed in many countries, one scale has come to dominate internationally, the Common European Framework of Reference for Languages (Council of Europe, 2001). This framework describes levels of language proficiency, originally in the languages of Europe, built on work carried out within the Council of Europe in the early 1970s. This work was intended to facilitate economic and cultural integration in Europe and in particular the transferability of educational credentials in languages as part of efforts to prepare a mobile workforce within Europe, an early, localized example of globalization. Although not a test itself, but rather a guiding curriculum and assessment framework, the Common European Framework of Reference, by specifying the terms by which communicative skill is to be understood, effectively determines the constructs of assessments related to it. Given that the construct of language tests is increasingly determined as a matter of policy in this way, the capacity of language testers to address issues of construct underrepresentation becomes extremely limited. Once constructs are enshrined in policy, they can only be changed by political means.

Specific testing programs, too, have been developed to implement policy. In educational settings, particular attention has focused on the impact of the U.S. government's No Child Left Behind Act of 2001, which involves standardized testing of learners at various key stages in schooling as a means of educational management and reform, particularly through the identification of weaknesses in the system in terms of results. One of the key areas that is tested is language. Given the linguistically very diverse populations of U.S. schools, particularly in urban areas with high levels of recent immigration, the administration of a single test for all learners regardless of linguistic background has meant that schools with high populations of children who are English language learners are at risk of performing poorly on such tests, with the punitive sanctions that result. The construct of such tests is not sensitive to the realities of bilingual development in children (see also Volume 3, Chapters 9, 10, and 17, this handbook, on school assessment, particularly in relation to culturally diverse school populations). At the international level, educational policy is increasingly being driven by results on the comparative international assessments of reading at age 15, the Programme for International Student Assessment, sponsored by the Organisation for Economic Co-operation and Development. For comparability to be achieved across educational systems conducted in different languages, the same texts, in

translation, must be used in each participating country. Hence, the texts must be of universal equivalent accessibility; no local cultural textual practices can be included in the texts chosen. For example, if a particular education system focuses heavily on literary or religious texts, this orientation is not reflected in the test. The search for texts that are universally acceptable means that functionalist, practical texts are heavily featured in the test. In fact, this selection is appropriate for the goal of the organization commissioning the Programme for International Student Assessment, which is hoping to provide measures of the readiness of the future workforce to participate in the globalized workplace. The impact of the results of the test on educational systems worldwide, but particularly in Europe, has been profound. The results are headline news in many countries, with the fate of education ministers frequently at stake. Documentation of the effects of the reading tests in different national settings is taking place (e.g., McNamara 2011b; Van Avermaet & Pulinx, 2010). Ironically, the impact of the test has been found to be positive in relation to minority languages in some settings (e.g., Lasagabaster, 2010, on Basque in the Spanish Basque Country). The unpredictability of the use and impact of language testing means that the idea that language testing practice is inherently ethical or unethical is too simplistic. In the light of this unpredictability, some have argued that language testers should focus their energies on making tests as fair as possible (see also Chapter 17, this volume), in the technical sense of reducing construct-irrelevant variance and construct underrepresentation, while leaving questions of the justice of tests (the legitimacy of their use) to others (for this distinction, see McNamara & Ryan, 2011). Davies (1997, 2004), Hamp-Lyons (1997), and Kunnan (2000) have accordingly limited the responsibilities of language testers in their discussions of the ethical responsibilities of language testers. Nevertheless, it could be argued that these discussions do not do justice to the complex issues raised by the policy-driven language testing practices referred to earlier.

CURRENT DEVELOPMENTS

Although performance tests of the type emerging from the communicative tradition are argued to have a positive influence on the teaching and learning leading up to the tests, because they focus on the real-world tasks for which the candidates need to be prepared (they have good "washback"; Cheng, Watanabe, & Curtis, 2004), they are complex and relatively expensive to develop and administer. A constant search exists for cheaper, more efficient methods capable of yielding a similar quality of information on relevant aspects of learner ability. The use of technology is transforming scoring of performances, and technological advances in the testing of speaking and listening are of particular interest and significance (Xi, 2010). Using highly artificial tasks such as reading aloud, simple sentence construction, or supplying opposites of given lexical items, new automated tests of speaking elicit from candidates a series of predictable utterances that are matched against a vast database to permit automatic computational analysis of features of performance such as phonetic quality and phonological and other fine-grained aspects of the realizations of the utterances. The resulting scores have been found to correlate sufficiently closely to those derived from a far more extensive oral proficiency interview to act as a useful and much cheaper proxy for it. The test, which is administered and scored automatically, can be taken by telephone at a time and place of the individual's choosing, takes no more than 10 minutes, costs relatively little, and provides an immediate score and report. The artificiality of the tasks and the underlying psycholinguistic construct clearly take the field back, in principle, to the tradition of performance assessment initiated by Oller (1979), although of course he did not foresee these precise developments. The construct in this automatic testing of speaking involves minute features of psycholinguistic processing, including reception and production. These processes underlying communication can be sampled and analyzed in a systematic way from performance on even artificial tasks and are then found to be predictive of measures of performance on more naturalistic tasks. Although the cost of administering and scoring such a test is a fraction of that of more naturalistic measures involving human judgment, the cost of establishing the analytic capacity underlying the test is enormous, which puts it beyond the reach of most

testing agencies, let alone individuals. The one or two organizations that own the capacity for such tests thus have a monopoly on their use.

Theories of learning are also changing and affecting language testing. Of particular interest here is the influence of neo-Vygotskian theory in focusing assessment as much on the potential for future performance as on current performance. Explorations in dynamic assessment (Lantolf & Poehner, 2008) are promising for guiding classroom-based assessment as an alternative to performance on standardized tests. The current interest in dynamic assessment is part of a movement to refocus work in language testing away from high-stakes language tests and onto assessment, which is more responsive to the educational setting and the needs of teachers and learners (Rea-Dickins, 2008). This change can be observed in the contrast between subsequent editions of the *Language Testing and Assessment* volume of the *Encyclopedia of Language and Education* (Clapham & Corson, 1997; Shohamy & Hornberger, 2008), with the more recent edition focusing extensively on assessment in a variety of educational settings.

A further development has been a questioning of the performance of the native speaker as the relevant reference point in second language performance, which has a long tradition in applied linguistics (Davies, 2008b). This challenge to the standing of the native speaker in relation to language tests is particularly true for English as a second or foreign language, which is used more among non-native speakers (where it acts as a lingua franca) than it is between native and nonnative speakers. Here the priorities in communication are cooperation, readiness to negotiate meaning when misunderstandings occur, and so on. The testing of English as a lingua franca (McNamara, 2011a; Seidlhofer, 2003) is in its infancy, but it will involve a rethinking of the construct and the development of more relevant assessment criteria. It is likely to have a considerable impact on English language testing and on language testing more broadly.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1–34. doi:10.1207/s15434311laq0201_1

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238–257. doi:10.1177/026553229501200206

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford, England: Oxford University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. doi:10.1037/0033-295X.111.4.1061

Brindley, G. (2008). Educational reforms and language testing. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment.* (2nd ed., pp. 365–378). Dordrecht, the Netherlands: Springer.

Brown, A. (2005). *Interviewer variability in language proficiency interviews.* Frankfurt, Germany: Peter Lang.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientation and test-taker performance on English for Academic Purposes speaking tasks* (TOEFL Monograph Series MS-29). Princeton, NJ: Educational Testing Service.

Cameron, R., & Williams, J. (1997). Senténce to ten cents: A case study of relevance and communicative success in nonnative–native speaker interactions in a medical setting. *Applied Linguistics, 18,* 415–445. doi:10.1093/applin/18.4.415

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1–47. doi:10.1093/applin/1.1.1

Carroll, J. B. (1954). *Notes on the measurement of achievement in foreign languages.* Unpublished mimeo.

Carroll, J. B. (1972). Fundamental considerations in testing for English language proficiency in foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (2nd ed., pp. 313–321). New York, NY: McGraw-Hill. (Original work published 1961)

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language.* New York, NY: Routledge.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods.* Mahwah, NJ: Erlbaum.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Clapham, C., & Corson, D. (Eds.). (1997). *Encyclopaedia of language and education: Vol. 7. Language testing*

and assessment. Dordrecht, the Netherlands: Kluwer Academic.

Clark, J. L. D., & Clifford, R. T. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques: Development, current status and needed research. *Studies in Second Language Acquisition, 10,* 129–147. doi:10.1017/S0272263100007270

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching and assessment*. Cambridge, England: Cambridge University Press.

Davies, A. (Ed.). (1997). Ethics in language testing [Special issue]. *Language Testing, 14*(3).

Davies, A. (Ed.). (2004). The ethics of language assessment [Special issue]. *Language Assessment Quarterly,* 1(2–3).

Davies, A. (2008a). *Assessing academic English: Testing English proficiency 1950–1989—The IELTS solution*. Cambridge, England: Cambridge University Press.

Davies, A. (2008b). The native speaker in applied linguistics. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 431–450). Oxford, England: Blackwell. doi:10.1002/9780470757000.ch17

de Saussure, F. (1983). *Course in general linguistics* (C. Bally & A. Sechehaye, Trans.). La Salle, IL: Open Court.

Eades, D. (2009). Testing the claims of asylum seekers: The role of language analysis. *Language Assessment Quarterly, 6*, 30–40. doi:10.1080/15434300802606523

Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives on integration regimes*. London, England: Continuum.

Hamp-Lyons, L. (1997). Ethics in language testing. In C. M. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 323–333). Dordrecht, the Netherlands: Kluwer Academic.

Hogan-Brun, G., Mar-Molinero, C., & Stevenson, P. (Eds.). (2009). *Discourses on language and integration: Critical perspectives on language testing regimes in Europe*. Amsterdam, the Netherlands: John Benjamins.

Hymes, D. H. (1972a). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 35–71). New York, NY: Holt, Rinehart & Winston.

Hymes, D. (1972b). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Harmondsworth, England: Penguin.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics, 32*, 23.1–23.17. doi:10.2104/aral0923

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14,* 47–84. doi:10.1177/026553229701400104

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—A longitudinal study. *Language Testing, 28*, 179–200. doi:10.1177/0265532210384252

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, England: Cambridge University Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, England: Longmans, Green.

Lantolf, J. P., & Poehner, M. E. (2008). Dynamic assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (2nd ed., pp. 273–284). Dordrecht, the Netherlands: Springer.

Lasagabaster, D. (2010, September). The linguistic side-effects of the PISA report in a bilingual context. In T. McNamara (Chair), *PISA, multilingualism and L1 language-in-education policy: A study of the impact of the Program for International Student Assessment (PISA) reading test in five national contexts*. Symposium conducted at the Sociolinguistics Symposium 18 (SS18), Southampton, England.

Long, M. (Ed.). (2005). *Second language needs analysis*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511667299

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*, 397–421. doi:10.1177/0265532209104668

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7,* 52–76. doi:10.1177/026553229000700105

McNamara, T. (1996). *Measuring second language performance*. London, England: Addison-Wesley Longman.

McNamara, T. (2005). 21st century Shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4,* 351–370. doi:10.1007/s10993-005-2886-0

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3,* 31–51. doi:10.1207/s15434311laq0301_3

McNamara, T. (2011a). Managing learning: Authority and language assessment. *Language Teaching, 4,* 500–515. doi:10.1017/S0261444811000073

McNamara, T. (2011b). Measuring deficit. In C. N. Candlin & J. Crichton (Eds.), *Discourses of deficit* (pp. 311–326). London, England: Palgrave Macmillan.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

McNamara, T., & Ryan, K. (2011). Fairness vs. justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8,* 161–178.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

No Child Left Behind Act of 2001, Pub. L. 107–110, 115 Stat. 1425 (2002).

Oller, J. W. (1979). *Language tests at school*. London, England: Longman.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12,* 217–237. doi:10.1177/026553229501200205

O'Loughlin, K. (2001). *Studies in language testing: Vol. 13. The equivalence of direct and semi-direct speaking tests* (M. Milanovic & C. Weir, Series Eds.). Cambridge, England: Cambridge University Press.

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16,* 9–13. doi:10.1111/j.1745-3992.1997.tb00586.x

Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (2nd ed., pp. 257–271). Dordrecht, the Netherlands: Springer.

Schegloff, E. A., Koshik, I., Jacoby, S., & Olsher, D. (2002). Conversation analysis and applied linguistics. *Annual Review of Applied Linguistics, 22*, 3–31. doi:10.1017/S0267190502000016

Seidlhofer, B. (2003). *A concept of international English and related issues: From "real English" to "realistic English."* Strasbourg, France: Council of Europe, Language Policy Division DG IV, Directorate of School, Out-of-School and Higher Education.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Pearson.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. New York, NY: Routledge.

Shohamy, E., & Hornberger, N. H. (Eds.). (2008). *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (2nd ed.). Dordrecht, the Netherlands: Springer.

Spolsky, B. (1978). Introduction: Linguists and language testers. In B. Spolsky (Ed.), *Advances in language testing series: Vol. 2. Approaches to language testing.* (pp. v–x). Arlington, VA: Center for Applied Linguistics.

Spolsky, B. (1996). The impact of the Army Specialized Training Program: A reconsideration. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 323–334). Oxford, England: Oxford University Press.

Taylor, L., & Wigglesworth, G. (Eds.). (2009). Pairwork in L2 assessment contexts [Special issue]. *Language Testing*, *26*(3).

Van Avermaet, P., & Pulinx, R. (2010, September). PISA—Flanders (Belgium). In T. McNamara (Chair), *PISA, multilingualism and L1 language-in-education policy: A study of the impact of the Program for International Student Assessment (PISA) reading test in five national contexts*. Symposium conducted at the Sociolinguistics Symposium 18 (SS18), Southampton, England.

Weir, C. J. (1983). The Associated Examining Board's Test in English for Academic Purposes: An exercise in content validation. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 147–153). London, England: Academic Press.

Xi, X. (Ed.). (2010). Automated scoring and feedback systems for language assessment and learning [Special issue]. *Language Testing*, *27*(3).

Young, R., & He, A. W. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam, the Netherlands: John Benjamins.

PART III

INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY

CHAPTER 22

ASSESSMENT IN INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY: AN OVERVIEW

John P. Campbell

The basic theme of this chapter is that the assessment enterprise in industrial and organizational (I/O) psychology is very broad, very complex, and very intense. The major underlying reason is that the world of work constitutes the major portion of almost everybody's adult life, over a long period of time. It *is* complicated. The major components of this complexity are the broad array of variables that must be assessed; the multidimensionality of virtually every one of them; the difficulties involved in developing specifications for such a vast array of variables; the wide variety of assessment methods; the intense interplay among science, research, and practice; and the critical value judgments that come into play. This chapter gives a structured overview of these issues, with particular reference to substantively modeling psychology's major variable domains and the attendant assessment issues that are raised. The conclusion is that substantive specifications for what psychologists are trying to assess are critically important, and I/O psychologists should not shortchange this requirement, no matter how much the marketplace seems to demand otherwise.

To be fair, the term *assessment* can take on different meanings. Perhaps its narrowest construction is as a multifactor evaluation of specific individuals in terms of their suitability for a specific course of action, such as selection, training, or promotion. However, if the full spectrum of research and practice concerning the applications of psychology to the world of work is considered, assessment becomes a much, much broader activity. This chapter takes the broadest perspective. It equates assessment with measurement and outlines a map of the assessment landscape. The landscape is described in terms of (a) an overall framework of relationships that describe what I/O psychology is about, (b) the range of assessment purposes that flow from this framework, (c) the range and complexity of the variables that require assessment, (d) the range and complexity of the assessment methods that can be used, and (e) the psychometric issues that permeate the assessment enterprise.

In the beginning were the independent variable and the dependent variable, a distinction that sounds sophomoric but is of fundamental importance and is often neglected. For example, when discussing the history of assessing leadership, a distinction is often made between trait models and behavioral models as though they were competing explanations (e.g., Hunt, 1999). However, the behavioral models (e.g., Bowers & Seashore, 1966) focus on leader performance—the dependent variable—and trait models focus on a particular set of performance determinants (e.g., cognitive ability, personality)—the independent variables. The dependent variable is the variable of real interest. It is the variable one wants to predict, enhance, or explain for various value-laden reasons. The independent variable has no intrinsic, or extrinsic, value. For example, knowing someone's general cognitive ability has no intrinsic value. It only has value because it predicts, or does not predict, something else that is of value (e.g., leadership performance). Similarly, independent variables such as training

DOI: 10.1037/14047-022
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

programs or motivational interventions have no value unless they can change something that is important (i.e., critical dependent variables).

DEPENDENT VARIABLE LANDSCAPE

So what then are the dependent variables of value that populate the I/O psychology landscape? Identifying the relevant set is indeed a value judgment, and the superordinate distinction is whether one takes the individual or the institutional (i.e., organizational) point of view (Cronbach & Gleser, 1965). That is, is it the values of the management that determine what dependent variables are important, or the values of the individual job holder? The management cares about the viability of the organization. Individuals care about their own viability. Sometimes their respective concerns overlap. For example, the management values high individual performance because it contributes to the goals of the organization. Individuals strive for high performance because it improves their standard of living, long-term financial security, or sense of self-worth. However, for the individual, higher and higher levels of performance may lose value because the effort required to achieve them detracts from other dependent variables, such as one's general life satisfaction.

Wherein lie the values of the researcher and scientist? One argument is that the researcher and scientist must choose between the values of the organization and the values of the individual. Once that choice is made, then the interests of the scientist focus on determining the best methods of assessment, given the purposes for which the information is to be used. An alternative argument is that the scientist does not make the value judgment. A dependent variable, such as individual performance, is modeled and measured for the purpose of studying its determinants. Such research can be used both by the organization to improve selection and by the individual to improve career planning. The intent here is not to settle such arguments but to make the point that value judgments permeate all choices of what to assess on the dependent variable side. It is also tempting to argue that values do not intrude on the independent variable side where the canons of psychometric theory preside, but obviously such is not the case, as discussed in a later section of the chapter. Those value judgments pertain to the consequences of the decisions made as a function of assessment of the independent variable. A partial taxonomy of the dependent variables in I/O psychology follows.

From the organization's point of view, the dependent variables are

- individual performance in a work role, including individual performance as a team member;
- voluntary turnover;
- team performance as a team, not as the aggregation of individual contributions;
- team viability (analogous to individual turnover);
- productivity (in the economist's sense) of (a) individuals, (b) teams, and (c) organizational units; and
- organizational unit effectiveness (i.e., the bottom line).

From the individual's point of view, the dependent variables are

- career and occupational achievement;
- satisfaction with the outcomes of working (which could include satisfaction with performance achievement);
- perceived (or experienced) fair treatment (e.g., distributive and procedural justice);
- frequency of injury from accidents; and
- overall health and well-being, including physical and mental health, perceived stress, and work–family conflict.

These two lists carry at least the following assumptions, qualifications, or both:

1. Organizations are not concerned about job satisfaction or subjective well-being as dependent variables, but only as independent variables that have implications for performance, productivity, effectiveness, or turnover.
2. Information pertaining to the determinants of performance may be used in a selection system, to benefit the organization, or in a career guidance system, to benefit the individual (e.g., using ability, personality, and interest assessment to plan educational or job search activities).

Similarly, training programs that produce higher skill levels can enhance individual performance for the benefit of the organization or enhance career options for individuals.

3. Fair and equitable treatment of individual employees and the level of individual health and well-being may be important dependent variables for the organization if they are incorporated as goals in the organization's ethical code or in a policy statement of corporate social responsibility, for which the management is then held responsible.

For the most part, I/O psychology does not operate from the individual point of view, even though several of its early pioneers did, for example, Donald Paterson or Walter van Dyke Bingham (cf. Koppes, Thayer, Vinchur, & Salas, 2007). At some point, vocational psychology (i.e., the individual point of view) became part of counseling psychology (Campbell, 2007; Meyer, 2007).

The dependent variable landscape is complex for assessment purposes, even as illustrated by the preceding simple lists. The complexity of assessment increases considerably when each of the general variables is modeled in terms of its major components. Consider each of the following.

Individual Performance

Before the mid-1980s, there was, relative to the assessment of individual performance, simply the "criterion problem" (J. T. Austin & Villanova, 1992), which was the problem of finding some existing and applicable indicator that could be construed as a measure (i.e., assessment) of individual performance (e.g., sales, number of pieces produced) while not worrying too much about the validity, reliability, deficiency, and contamination of the indicators. Since then, much has happened regarding how performance is defined and how its latent structure is modeled.

In brief, the consensus is that *individual performance* is best defined as consisting of the actions people engage in at work that are directed at achieving the organization's goals and that can be scaled in terms of how much they contribute to said goals. For example, sometimes it takes a great deal of covert thinking before the individual does something. Performance is the action, not the thinking that preceded the action, and someone must identify those actions that are relevant to the organization's goals and those that are not. For those that are (i.e., performance), the level of proficiency with which the individual performs them must be scaled. Both the judgment of relevance and the judgment of level of proficiency depend on a specification of the organization's important substantive goals, not content-free goals such as "make a profit."

Nothing in this definition requires that a set of performance actions be circumscribed by the term *job* or that they remain static over a significant length of time. Neither does it require that the goals of an organization remain fixed or that a particular management cadre be responsible for determining the organization's goals (also know as *vision*). However, for performance assessment to take place, the major operative goals of the organization, within some meaningful time frame, must be known, and the methods by which individual actions are judged to be goal relevant, and scaled in terms of what represents high and low proficiency, must be legitimized by the stakeholders empowered to do so by the organization's charter. Otherwise, there is no organization. This is as true for a family as it is for a corporation.

This definition creates a distinction between performance, as defined earlier, and the outcomes of performance (e.g., sales level, incurred costs) that are not solely determined by the performance of a particular individual, even one of its top executives. If these outcome indicators represent the goals of the organization, then individual performance should certainly be related to them. If not, the specifications for individual performance are wrong and need changing or, conversely, the organization is pursuing the wrong goals. If the variability in an outcome indicator is totally under the individual's control, then it is a measure of performance.

Given an apparent consensus on this definition of performance, considerable effort has been devoted to specifying the dimensionality of performance, in the context of the latent structure of the performance actions required by a particular occupation, job, position, or work role (see Bartram,

2005; Borman & Brush, 1993; Borman & Motowidlo, 1993; Campbell, McCloy, Oppler, & Sager, 1993; Griffin, Neal, & Parker, 2007; Murphy, 1989a; Organ, 1988; Yukl, Gordon, & Taber, 2002). These models have become known as performance models, and they seem to offer differing specifications for what constitutes the nature of performance as a construct. However, the argument here is that correspondence is virtually total.

Campbell (2012) has integrated all past and current specifications of the dimensional structure of the dependent variable, individual performance, including those dealing with leadership and management performance, and the result is summarized in the eight basic factors discussed in the next section.

Orthogonality is not asserted or implied, but content distinctions that have different implications for selection, training, and organizational outcomes certainly are. Although scores on the different dimensions may be added together for a specific measurement purpose, it is not possible to provide a substantive specification for a "general" factor. Whether dimensions can be as general as contextual performance or citizenship behavior is also problematic.

Basic factors. The basic substantive factors of individual performance in a work role (which are not synonymous with Campbell et al., 1993) are asserted to be the following.

Factor 1: Technical Performance. All models acknowledge that virtually all jobs or work roles have technical performance requirements. Such requirements can vary by substantive area (driving a vehicle vs. analyzing data) and by level of complexity or difficulty within area (driving a taxi vs. driving a jetliner; tabulating sales frequencies vs. modeling institutional investment strategies). Technical performance is not to be confused with task performance. A task is simply one possible unit of description that could be used for any performance dimension.

The subfactors for this dimension are obviously numerous, and the domain could be parsed into wide or narrow slices. The Occupational Information Network (O*NET; Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999) is based on the U.S. Department of Labor's Standard Occupational Classification structure, which currently uses 821 occupations for describing the major distinctions in technical task content across the entire labor force, and the 821 occupations are further aggregated into three higher order levels consisting of 449, 96, and 23 occupational clusters, respectively. The managers of O*NET have interestingly divided some of the Standard Occupational Classifications into narrower slices to better suit user needs and have also added new and emerging occupations such that O*NET 14.0 collected data on 965 occupations. The number will grow in the future (Tippins & Hilton, 2010). Potentially, at least, an occupational classification based on technical task content could be used to archive I/O psychology assessment data on individual work-role performance, end-of-training performance, or predicted performance.

Factor 2: Communication. The Campbell et al. (1993) model is the only one that isolates communication as a separate dimension, but it appears as a subfactor in virtually all others. *Communication* refers to the proficiency with which one conveys information that is clear, understandable, and well organized. It is defined as being independent of subject matter expertise. The two major subfactors are oral and written communication.

Factor 3: Initiative, Persistence, and Effort. This factor emerged from the contextual performance and management performance literatures as well as the organizational citizenship behavior literature in which it was referred to as Individual Initiative. To make this factor conform to the definition of performance used here it must be composed of observable actions. Consequently, it is typically specified in terms of extra hours, voluntarily taking on additional tasks, going beyond prescribed responsibilities, working under extreme or adverse conditions, and so forth.

Factor 4: Counterproductive Work Behavior. *Counterproductive Work Behavior* (CWB), as it has come to be called, refers to a category of individual actions or behaviors that have negative implications for accomplishment of the organization's goals (see Chapter 35, this volume, for additional information on this area).

The current literature does not speak with one voice regarding the meaning of CWB, but the

specifications generally circumscribe actions that are intentional, that violate or deviate from prescribed norms, and that have a negative effect on the individual's contribution to the goals of the unit or organization. Descriptions of this domain are provided by Gruys and Sackett (2003) and Robinson and Bennett (1995). The general agreement seems to be that two major subfactors exist (e.g., see R. J. Bennett & Robinson, 2000; Berry, Ones, & Sackett, 2007; Dalal, 2005) distinguished by deviant behaviors directed at the organization (theft, sabotage, falsifying information, malingering) and behaviors directed at individuals, including the self (e.g., physical attacks, verbal abuse, sexual harassment, drug and alcohol abuse). Although not yet fully substantiated by research, it seems reasonable to also expect an approach–avoidance, or moving toward versus moving away, distinction for both organizational deviance and individual deviance. That is, the CWBs dealing with organizational deviance seem to be divided between aggressively destroying or misusing resources versus avoiding or withdrawing from the responsibilities of the work role. Similarly, CWBs directed at individuals seem to be divided between aggressive actions that are directed at other people and destructive actions directed at the self, such as alcohol and drug abuse and neglect of safety precautions. The approach–avoidance distinction is a recurring one in the study of motivation (Elliot & Thrash, 2002; Gable, Reis, & Elliot, 2003) and of personality (Watson & Clark, 1993), including a major two-factor model of psychopathology (Markon, Krueger, & Watson, 2005). It is also suggested in a study of CWB by Marcus, Schuler, Quell, and Humpfner (2002).

A major issue in the CWB literature is whether its principal subfactors are simply the extreme negative end of other performance factors or whether they are independent constructs. The evidence currently available (Berry et al., 2007; Dalal, 2005; Kelloway, Loughlin, Barling, & Nault, 2002; Miles, Borman, Spector, & Fox, 2002; Ones & Viswesvaran, 2003; Spector, Bauer, & Fox, 2010) has suggested that CWBs are not simply the negative side of other performance components. Low scores on other performance dimensions could result from a lack of knowledge or skill, but low scores on CWB reflect intentional deviance and are dispositional in origin.

Factor 5: Supervisory, Manager, Executive (i.e., hierarchal) Leadership. This factor refers to leadership performance in a hierarchical relationship. The substantive content, as specified by the leadership research literature, is most parsimoniously described by the six leadership factors listed in Exhibit 22.1 (Campbell, 2012). The parsimony results from the remarkable convergence of the literature, as detailed in Campbell (2012), from the Ohio State and Michigan studies through the contingency theories of Fielder, House, Vroom, and Yetton to the current emphasis on being charismatic and

Exhibit 22.1
Six Basic Factors Making Up Leadership Performance

1. *Consideration, support, person centered:* Providing recognition and encouragement, being supportive when under stress, giving constructive feedback, helping others with difficult tasks, building networks with and among others
2. *Initiating structure, guiding, directing:* Providing task assignments; explaining work methods; clarifying work roles; providing tools, critical knowledge, and technical support
3. *Goal emphasis:* Encouraging enthusiasm and commitment for the group's or organization's goals, emphasizing the important missions to be accomplished
4. *Empowerment, facilitation:* Delegating authority and responsibilities to others, encouraging participation, allowing discretion in decision making
5. *Training, coaching:* One-on-one coaching and instruction regarding how to accomplish job tasks, how to interact with other people, and how to deal with obstacles and constraints
6. *Serving as a model:* Models appropriate behavior regarding interacting with others, acting unselfishly, working under adverse conditions, reacting to crisis or stress, working to achieve goals, showing confidence and enthusiasm, and exhibiting principled and ethical behavior.

Note. From *Oxford Handbook of Industrial and Organizational Psychology* (p. 173), by S. Kozlowski (Ed.), 2012, New York, NY: Oxford University Press. Copyright 2012 by Oxford University Press. Adapted with permission.

transformational, leading the team, and operating in highly complex and dynamic environments. In conversations about leadership, the emphasis may be on leader performance, as defined here, or it may be on the outcomes of leader actions (e.g., follower satisfaction, unit profitability), on the determinants (predictors) of leadership performance, or on the contextual influences on leader performance or performance outcomes. However, when describing or assessing leadership performance (as defined here), the specifications are always in terms of one or more of these six factors. The relative emphasis may be different, and different models may hypothesize different paths from leader performance to leader effectiveness, which for some people may be the interesting part, but the literature's characterization of leader performance itself seems to always be within the boundaries of these six subfactors.

Similarly, the six subfactors circumscribe hierarchical leadership performance at all organizational levels. However, the relative emphasis on the factors may change at higher organizational levels, and the specific actions within each subfactor may also receive differential emphasis.

Factor 6: Management Performance (hierarchical). Within a hierarchical organization, this factor includes those actions that deal with obtaining, preserving, and allocating the organization's resources to best achieve its goals. The major subfactors of management performance are given in Exhibit 22.2 (Campbell, 2012). The major distinction between leadership performance and management performance, which not everybody agrees on, is that the leadership dimensions involve interpersonal influence. The management dimensions do not. As it was for the components of leadership, there may be considerably different emphases on the management performance subfactors across work roles and also as a function of the type of organization, organizational level, changes in the situational context, changes in organization goals, and so forth. Also, nothing in the leadership–management distinction implies two separate jobs or work roles. They coexist.

Factor 7: Peer–Team Member Leadership Performance. The content of this factor is parallel to the actions that make up hierarchical leadership (see Factor 5). The defining characteristic is that these actions are in the context of peer or team member interrelationships, and the peer–team relationships in question can be at any organizational level (e.g., production teams vs. management teams). That is, the team may consist of nonsupervisory roles or a team of unit managers.

Factor 8: Team Member–Peer Management Performance. A defining characteristic of the

Exhibit 22.2
Eight Basic Factors of Management Performance

1. *Decision making, problem solving, and strategic innovation:* Making sound and timely decisions about major goals and strategies. Includes gathering information from both inside and outside the organization, staying connected to important information sources, forecasting future trends, and formulating strategic and innovative goals to take advantage of them
2. *Goal setting, planning, organizing, and budgeting:* Formulating operative goals; determining how to use personnel and resources (financial, technical, logistical) to accomplish goals; anticipating potential problems; estimating costs
3. *Coordination:* Actively coordinating the work of two or more units or the work of several work groups within a unit; scheduling operations; includes negotiating and cooperating with other units
4. *Monitoring unit effectiveness:* Evaluating progress and effectiveness of units against goals: monitoring costs and resource consumption
5. *External representation:* Representing the organization to those not in the organization (e.g., customers, clients, government agencies, nongovernment organizations, the public); maintaining a positive organizational image; serving the community; answering questions and complaints from outside the organization
6. *Staffing:* Procuring and providing for the development of human resources; not one-on-one coaching, training, or guidance, but providing the human resources that the organization or unit needs
7. *Administration:* Performing day-to-day administrative tasks, keeping accurate records, documenting actions; analyzing routine information and making information available in a timely manner
8. *Commitment and compliance:* Compliance with the policies, procedures, rules, and regulations of the organization; full commitment to orders and directives, together with loyal constructive criticism of organizational policies and actions

Note. From *Oxford Handbook of Industrial and Organizational Psychology* (p. 173), by S. Kozlowski (Ed.), 2012, New York, NY: Oxford University Press. Copyright 2012 by Oxford University Press. Adapted with permission.

high-performance work team (e.g., Goodman, Devadas, & Griffith-Hughson, 1988) is that team members perform many of the management functions shown in Exhibit 22.2. For example, the team member performance factors identified in a critical incident study by Olson (2000) that are not accounted for by the technical performance factors, or the peer leadership factors, concern such management functions as planning and problem solving, determining within-team coordination requirements and workload balance, and monitoring team performance. In addition, the contextual performance and organizational citizenship behavior literatures have both strongly indicated that representing the unit or organization to external stakeholders and exhibiting commitment to and compliance with the policies and procedures of the organization are critical performance factors at any organizational level. Consequently, to a greater extent than most researchers realize or acknowledge, important elements of management performance exist in the peer or team context as well as in the hierarchical (i.e., management–subordinate) setting.

Again, these eight factors are intended to be an integrative synthesis of what the literature has suggested are the principal dimensions of performance in a work role. They are meant to encompass all previous work on individual performance modeling, team member performance, and leadership and management. Even though the different streams of literature may use somewhat different words for essentially the same performance actions, great consistency exists across the different sources.

Performance dynamics. The latent structure just summarized has direct implications for the content of performance assessments. However, it does not speak to whether an individual's level of performance is stable over time or whether it changes. Assessment of performance dynamics must deal with additional complexities. One source of such dynamics is that performance requirements of the work role itself change over time, which can occur because of changes in (a) the substantive content of the requirements, (b) the level of performance expected, (c) the conditions under which a particular level of performance is expected, or (d) some combination of these. Individuals can also change. Much of I/O psychology research and practice deals with planned interventions designed to enhance the individual knowledge, skill, and motivational determinants of performance, such as training and development, goal setting, feedback, rewards of various kinds, better supervision, and so forth. Such interventions, with performance requirements held constant, could increase the group mean, have differential effects across people, or both. The performance changes produced can be sizable (e.g., Carlson, 1997; Katzell & Guzzo,1983; Locke & Latham, 2002).

Interventions designed to enhance individual performance determinants can also be implemented by the individual's own processes of self-management and regulation (Kanfer, Chen, & Pritchard, 2008; Lord, Diefendorff, Schmidt, & Hall, 2010), and the effectiveness of these self-regulation processes could vary widely across people. In addition, if they have the latitude to do so, individuals could conduct their own job redesign (i.e., change the substantive content of their work role) to better utilize their knowledge and skills and increase the effort they are willing to spend. Academics are fond of doing that.

As noted by Sonnentag and Frese (2012), individual performance can also change simply as a function of the passage of time. Of course, time is a surrogate for such things as practice and experience, the aging process, or changes in emotional states (Beal, Weiss, Barros, & MacDermid, 2005).

Most likely, for any given individual over any given period of time, many of these sources of performance change can be operating simultaneously. Performance dynamics are complex, and attempts to model the complexity have taken many forms. For example, there could be characteristic growth curves for occupations (e.g., Murphy, 1989b), differential growth curves across individuals (Hofmann, Jacobs, & Gerras, 1992; Ployhart & Hakel, 1998; Stewart & Nandkeolyar, 2006; Zyphur, Chaturvedi, & Arvey, 2008), both linear and nonlinear components for growth curves (Deadrick, Bennett, & Russell, 1997; Reb & Cropanzano, 2007; Sturman, 2003), and cyclical changes resulting from a number of self-regulatory mechanisms (Lord et al., 2010).

Empirical demonstrations of each of these have been established.

Adapting to dynamics. Adaptability can be viewed either as a characteristic of performance itself (i.e., a category of performance actions), as did Hesketh and Neal (1999), or as a property of the individual (i.e., as a determinant of performance). Ployhart and Bliese (2006) presented a thorough discussion of this issue and argued that it is more useful to model (i.e., identify the characteristics of) the adaptive individual than it is to propose adaptability as a distinct content dimension of performance. One reason is that the general definition of adaptability is not content domain specific, and providing specifications for adaptability as a distinct performance dimension has been difficult (e.g., see Pulakos, Arad, Donovan, & Plamondon, 2000).

Domain-specific dynamics. In sum, it can be taken as a given that work-role performance requirements change over time, sometimes over very short periods of time, and that individuals change (i.e., adapt) to meet them. Individuals can also change in anticipation of changes in performance requirements. Many interventions (e.g., training, goal setting, reward systems) have been developed to help individuals adapt to changing performance requirements. Individuals can also actively engage in their own self-management to develop additional knowledge and skill and to regulate the direction and intensity of their effort. If the freedom to do so exists, they can even proactively change their own performance responsibilities, or at least their relative emphases, so as to better use their own knowledge and skill or to better accomplish unit goals. Even if performance requirements remain relatively constant, individual performance can change over time as the result of practice, feedback, increasing experience, cognitive and physical changes resulting from aging, or even fluctuation in affect or subjective well-being.

As a result of all this, one might ask what implications performance dynamics and individual adaptability have for substantive models of individual work performance. This question is not the right question. A more appropriate question is, "What are the implications of substantive models of performance for the assessment of performance dynamics and individual adaptability?" The argument here is that although the latent dimensions of performance may be interdependent (e.g., higher technical performance could enhance leadership), the assessment of performance change must be linked to the individual performance dimensions. That is, the nature of performance changes may be different for different dimensions.

Summary. Why devote so much space to the basic modeling of individual performance in what is supposed to be an overview of assessment in I/O psychology? There are two reasons. First, individual performance is I/O psychology's most important dependent variable. Second, considering the assessment of individual performance raises some very fundamental issues that are relevant for the assessment of virtually all other variables, both dependent and independent. For example, what is the most useful specification for the latent structure? To what extent is the "most useful specification" a function of value judgments? Judgments by whom? Aside from conventional considerations of reliability, are the latent variables "dynamic"? What is the expected nature of the within-person variation? All of these issues have implications for the choice of assessment methods and for the purposes for which specific assessments are used.

Performance Assessment

The assessment of individual work-role performance may be I/O psychology's most difficult assessment requirement. J. T. Austin and Villanova (1992) provided ample documentation of the problem. Archival objective measures are few and far between and frequently suffer from contamination. Ratings, although they do yield meaningful assessments (W. Bennett, Lance, & Woehr, 2006; Conway & Huffcut, 1997), tend to suffer from low reliability, method variance, contamination, and the possible intrusion of implicit models of performance held by the raters that do not correspond to the stated specifications of the assessment procedure (Borman, 1987; Conway, 1998). Alternatives to ratings have been methods such as performance in a simulator, performance on various forms of job samples (Campbell & Knapp, 2001), and using various indicators of goal

accomplishment when goals are specified such that accomplishing them is virtually under the individual's total control (Pulakos & O'Leary, 2010).

In addition to these considerations, taking account of the purpose of assessment is also critical. The three major reasons for assessing performance are (a) for research purposes that have no high-stakes consequences; (b) for developmental purposes that carry the assurance that low scores do not carry negative consequences; and (c) for high-stakes appraisal situations such as promotion, compensation, termination, and so forth. Most likely, different assessment methods would be appropriate for each. Also, depending on which of the three is operative, the same assessment procedure could produce different assessments. For example, raters could be trying to satisfy different goals when doing operational performance appraisals versus providing ratings for research purposes only. Murphy and Cleveland (1995) discussed these issues at some length. The overall moral is that the measurement purposes must never be confused.

Team Performance

Research, theory, and professional discussion regarding team effectiveness, team performance, the determinants of team performance, and the processes by which the determinants (independent variables) affect team performance (dependent variables) has expanded exponentially over the past 20 years (e.g., Ilgen, Hollenbeck, Johnson, & Jundt, 2005; Kozlowski & Ilgen, 2006: Mathieu, Maynard, Rapp, & Gilson, 2008). However, most of the attention is given to the determinants of team performance and effectiveness and to the processes by which they have their effects. Modeling team performance itself for purposes of guiding assessment has received relatively little attention.

The dominant model is still that articulated by Hackman (1992), that is, that three major factors of group–team performance exist (as distinct from individual performance):

1. The first factor is the degree to which it accomplishes its major substantive task goals. This factor is analogous to the technical factor for individual performance. No taxonomy of team goals exists, but it could include such things as meeting production goals, producing solutions to specific problems, developing policy, creating designs, modeling resource allocation decisions, and so forth.
2. The second factor is the degree to which team members feel rewarded by, or satisfied with, their role and committed to the team's goals so that they continue to commit effort toward team goal accomplishment. This factor is analogous to the effort–initiative factor in individual performance.
3. The third factor is the degree to which the team improves its resources, skills, and coordination over time.

By implication, assessment of team performance would involve assessment of these three factors. The last two factors are sometimes combined into a higher order factor referred to as *team viability,* or the team's capability to maintain its technical performance over time.

Unit and Organizational Effectiveness

Organizations, and organizational units, do have a bottom line. That is, by some set of value judgments, a set of outcomes is identified that the organization or unit wants to maximize, optimize, or at least maintain at certain levels, such as quantity or quality of output (be it goods or services), sales, revenue, costs, earnings, return on investment, stock price, asset values, and so forth. The outcomes deemed important are a management choice, and choices can vary across organizations and across time within organizations. For an educational organization, the outcome could be number of students, graduation rates, time to degree, mean SAT or GRE scores for the student body, prestige of postgraduation job placements, and so forth. Again, by definition, the level and variation of such outcomes is the result of multiple determinants, in addition to individual performance. Although the term *organizational effectiveness* is used frequently in the I/O literature relative to both research and practice, attempts to model organizational or unit effectiveness for purposes of assessment have been sparse. An early taxonomy was developed by Campbell (1977), which was given a three-dimensional higher

order structure by Quinn and Rohrbaugh (1983) and Cameron and Quinn (1999).

Productivity

Productivity, particularly with regard to its assessment, is a frequently misused term in I/O psychology. Its origins are in the economics of the firm, where it refers to the ratio of the value of output (i.e., effectiveness) to the costs of achieving that level of output. Holding output constant, productivity increases as the costs associated with achieving that level of output decrease. It is possible to talk about the productivity of capital, the productivity of technology, and the productivity of labor, which are usually indexed by the value of output divided by the cost of the labor hours needed to produce it. For the productivity of labor, it would be possible to consider individual productivity, team productivity, or organizational productivity. Assessment of individual productivity would be a bit tricky, but it must be specified as the ratio of performance level (on each major dimension) to the cost of reaching that level (on each major dimension). Costs could be reflected by number of hours needed or wage rates. For example, terminating high wage-rate employees and hiring cheaper (younger?) individuals who can do the same thing would increase individual productivity.

Turnover

Turnover refers to the act of leaving an organization. Turnover can be voluntary or involuntary, as when an individual is terminated by the organization. Both voluntary turnover and involuntary termination can be good or bad depending on the circumstances. Depending on the work role, turnover could also vary as a function of determinants that operate at various times (e.g., variation in turnover could occur as a function of the initial socialization process, early vs. late promotions, vesting of retirement benefits).

For assessment purposes, great benefit would result if a latent structure for turnover could be specified in terms of the substantive reasons individuals leave. The beginnings of such a latent structure can be found in the integrative reviews of turnover research by Griffeth, Hom, and Gaertner (2000), Mitchell and Lee (2001), and Maertz and Campion (2004).

DEPENDENT VARIABLE ASSESSMENT FROM THE INDIVIDUAL'S POINT OF VIEW

Again, the defining characteristic is that higher scores on such variables are of value to the individual for his or her own sake. They are not of value because they correlate with or predict something else that is of value. Consequently, what is a dependent variable for the individual could be an independent variable for the organization.

Job Satisfaction

One taxonomy of such dependent variables valued by the individual is represented by the 20 dimensions assessed by the Minnesota Importance Questionnaire (Dawis & Lofquist, 1984), which are listed in Exhibit 22.3.

Within the theory of work adjustment (Dawis, Dohm, Lofquist, Chartrand, & Due, 1987; Dawis & Lofquist, 1984), the variables in Exhibit 22.3 are assessed in different ways for different reasons. The Occupational Reinforcer Pattern is a rating by supervisors or managers of the extent to which a particular work role provides outcomes representing each of the variables. The Minnesota Importance Questionnaire is a self-rating by the individual of the importance of being able to experience high levels of each of the 20 dimensions. The Minnesota Satisfaction Questionnaire is a self-rating of the degree to which the individual is satisfied with the level of each variable that he or she is currently experiencing. According to the theory of work adjustment, overall work satisfaction should be a function of the degree to which the work-role characteristics judged to be important by the individual are indeed provided by the work role, or job.

Exhibit 22.3 represents the literature's most finely differentiated portrayal of the latent structure of what individuals want from work. There are other portrayals. For example, a long time ago, Herzberg (1959) grouped 16 outcomes obtained via a critical incident procedure (he called it *story-telling*) into two higher order factors variously called *motivators and hygienes* or *intrinsic and extrinsic*. The Job

Exhibit 22.3
The 20 First-Level Job Outcomes Incorporated in Dawis and Lofquist's (1984) Minnesota Theory of Work and Adjustment

1. *Ability utilization:* The chance to do things that make use of one's abilities
2. *Achievement:* Obtaining a feeling of accomplishment and achievement from work
3. *Activity:* Being able to keep busy all the time, freedom from boredom
4. *Advancement:* Having realistic chances for promotion and advancement
5. *Authority:* Being given the opportunity to direct the work of others
6. *Company policies and practices:* Company policies and practices that are useful, fair, and well thought out
7. *Compensation:* Compensation that is fair, equitable, and sufficient for the work being done
8. *Coworkers:* Good interpersonal relationships among coworkers
9. *Creativity:* The opportunity to innovate and try out new ways of doing things in one's job
10. *Independence:* The chance to work without constant and close supervision
11. *Moral values:* Working does not require being unethical or going against one's conscience
12. *Recognition:* Receiving praise and recognition for doing a good job
13. *Responsibility:* The freedom to use one's own judgment
14. *Security:* Not having to worry about losing one's job
15. *Social service:* Opportunities to do things for other people as a function of being in a particular work role
16. *Social status:* The opportunity to be somebody in the community, as a function of working in a particular job and organization
17. *Supervision—human relations:* The respect and consideration shown by one's manager or supervisor
18. *Supervision—technical:* Having a manager or supervisor who is technically competent and makes good decisions
19. *Variety:* Having a job that incorporates a variety of things to do
20. *Working conditions:* Having working conditions that are clean, safe, and comfortable

Note. From *Oxford Handbook of Industrial and Organizational Psychology* (p. 173), by S. Kozlowski (Ed.), 2012, New York, NY: Oxford University Press. Copyright 2012 by Oxford University Press. Adapted with permission.

Descriptive Index (Smith, Kendall, & Hulin, 1969) focuses on five factors: the nature of the work itself; the characteristics of pay; the characteristics of supervision; the nature of promotion opportunities; and the characteristics of one's coworkers. There have also been several measures of overall, or general, job satisfaction (e.g., Hoppock, 1935; Kunin, 1955), which might use one item or several items.

Job satisfaction is a complex construct, and assessment issues revolve around the number of latent factors; the nature of the general factor; whether the sum of the parts (i.e., adding factor scores) captures all the variance in a rating of overall satisfaction; the dynamics of within-person variation; whether the frame of reference should be a description of the individual's state, an evaluation of that state, or the affective response to the evaluation; and how levels of satisfaction should be scaled (e.g., see Hulin & Judge, 2003). Assessment must deal with all of these issues.

It is instructive, or at least interesting, to compare the 20 job characteristics listed in Exhibit 22.3 with other individual work outcomes that the list does not seem to include but that have received important research or assessment attention. Examples follow.

Justice

A considerable literature exists on distributive and procedural justice (Colquitt, 2001; Colquitt, Conlon, Wesson, Porter, & Ng, 2001) that could be viewed as subfactors of Outcome 6 in Exhibit 22.3. *Distributive justice* refers to an individual's self-assessment of how well he or she is being rewarded by the organization. *Procedural justice* refers to the individual's assessment of the relative fairness of the organization's procedure for managing and dispensing rewards. A meta-analysis by Crede (2006) showed perceptions of procedural justice to have a somewhat higher mean correlation with overall job

satisfaction than did distributive justice (.56 vs. .62) when correlations were corrected for artifacts.

Overall Well-Being

Several dependent variables in the workplace, from the individual's point of view, go beyond job satisfaction and perceived distributive and procedural justice to include additional facets of overall well-being, such as the following:

- *Physical health:* In terms of its relationship to work roles, physical health is most often talked about in terms of a safe physical environment (Tetrick, Perrewé, & Griffin, 2010), that is, protections from environmental hazards, effective safety procedures, manageable physical demands, and available preventive care for potential illness. Assessment could involve the independent measurement of such factors or the individual's perception of them.
- *Mental and psychological health.* Although positive psychological health associated with working is a valued outcome from the individual point of view, it presents assessment complications. After controlling for basic personality characteristics, the framework proposed by Warr (1994) could be adopted that would then seek to assess (a) the individual's level of happiness or unhappiness, (b) relative feelings of comfort versus anxiety, and (c) feelings of depression versus enthusiasm. Lurking in the background is the research on set points (e.g., Lykken, 1999), which has argued that individuals have a characteristic level of happiness or well-being that determines much of the variance in their reactions to the work environment on these dimensions.
- *Work–family conflict.* This literature is growing, and the implication is that individuals value a work situation that does not produce undue conflict with family life or nonwork relationships. The determinants of work–family conflict are many and varied, and several models have been offered relating the determinants to work–family conflict (e.g., J. E. Edwards & Rothbard, 2000; Greenhaus & Powell, 2006; Grzywacz & Carlson, 2007). Some of the issues are whether work interferes with family or vice versa; whether the goals of the family and the goals of the individual at work are different; and the influence of gender (e.g., whether the man or woman stays home). The touchstone for assessment of the dependent variable is defining high scores as the perception (by the job holder) that work and family demands are in balance. That is, work demands do not degrade family goals, and family demands do not degrade individual work goals. Consequently, assessment should take into account how well the two sets of goals are aligned, and they may not be weighted equally (e.g., for economic reasons). Regardless of the relative weights, Cleveland and Colella (2010) made a strong argument for why both sets of goals strongly influence work–family conflict assessments.
- *Work-related stress.* The study of work stress has generated a very large literature (Sonnentag & Frese, 2003), and work stress is frequently offered as an important criterion variable because of the high frequencies with which it is reported (Harnois & Gabriel, 2000; Levi & Lunde-Jensen, 1996; National Institute for Occupational Safety and Health, 1999). *Stress* can be defined as a set of physiological, behavioral, or psychological responses to demands (work, family, or environmental) that are perceived to be challenging or threatening (Neuman, 2004). Assessment of individual stress levels is a more complex enterprise than assessment of job satisfaction, mental or physical health, or work–family conflict. The measurement operations could be physiological (e.g., cortisol levels in the blood), behavioral (e.g., absenteeism), psychological (depression), or perceptual (e.g., self-descriptions of stress levels), and the construct validity of any one of them is not assured given the complexities of modeling stress as a construct.

  A somewhat overly simplistic model of stress as a criterion would be that the work–family situation incorporates potential stressors. Whether a potential stressor (e.g., a new project deadline) leads to a stress reaction is a function of how it is evaluated by the individual. For some, the new deadline might be threatening (e.g., it increases the probability of a debilitating failure or makes

it difficult to care for a sick child). For others, it is merely an interesting challenge that will be fun to tackle. If potential stressors are evaluated as threatening, stress levels go up unless the individual has the resources to cope with them (Hobfoll, 1998). The Selye (1975) principle of optimum stress levels says that individuals need a certain amount of perceived stress to be optimally activated (Cooper, Dewe, & Driscoll, 2001). Similar models have been offered by Robert and Hockey (1997) and Warr (1987). However, if stress is too high, several counterproductive outcomes (labeled *strains*) can occur. These outcomes can be physical (fatigue, headaches), behavioral (reduced performance), or psychological (anxiety, sleep impairment). Consequently, assessment must choose among alternative measurement operations, must deal with the appraisal component (i.e., is a potential stressor actually a stressor?), and must make a case for the construct validity of the assessment of strains.

Individual Perspective: A Summary Comment

Job satisfaction, distributed and procedural justice, physical health, mental and psychological health, work–family conflict, stress, or simply evaluation of overall well-being have been discussed as dependent variables in the work setting that are important to individuals. That is, most people value being satisfied with their work, being physically and psychologically healthy, achieving a work life–non-work-life balance, and experiencing optimal stress levels. However, in the I/O psychology literature, these variables are usually not discussed as ends in themselves, but as independent variables that have an effect on the organization's bottom line (Cleveland & Colella, 2010; Tetrick et al., 2010). Depending on which perspective is chosen, the purpose of assessment is different, and the choice of assessment methods may differ as well.

INDEPENDENT VARIABLE LANDSCAPE

Compared with the dependent variable domain, the independent variable domain is a lush and verdant landscape—and much more intensely researched and assessed. It has also been well discussed by others and is the subject of many recent handbooks (Farr & Tippins, 2010; Scott & Reynolds, 2010; Zedeck, 2010). What follows is a brief outline primarily for the purpose of making certain distinctions that are discussed less often. As might be expected, the outline follows Campbell et al. (1993), Campbell and Kuncel (2001), and Campbell (2012).

The Campbell et al. (1993) model of performance posited two general kinds of performance determinants: direct and indirect. That is, individual differences in performance (either between or within) are a direct function of the current levels of performance-related knowledge and performance-related skills. There are different kinds of knowledge (e.g., facts, procedures) and different kinds of skills (e.g., cognitive, physical, psychomotor, expressive). The critical factor is that they are the real-time knowledge and skills determinants of performance. The only other direct determinants are motivational and are represented by three choices: (a) where to direct effort, (b) at what levels, and (c) for how long. All other performance determinants must exercise their effects by changing one or more of the direct determinants. It follows that a diagnosis of the direct causes of low or high performance must assess knowledge, skill levels, and choice behaviors that are specific to the work role's performance requirements in real time. For example, reading skill as a direct determinant refers to how well the individual reads the material required by the job in the work setting. Reading skill (ability?) as measured by the SAT is an indirect determinant. A multitude of indirect determinants of knowledge, skills, and choice behaviors exists, and a brief outline follows.

Traits: Abilities

The individual differences tradition in psychology in general, and I/O psychology in particular, has devoted much attention to the assessment of individual characteristics that are relatively stable over the adult working years. Assessments of such characteristics are used to predict future performance for selection and promotion purposes, predict who will benefit from specific training or development experiences, predict performance failures, provide the individual profiles needed to determine person–job

or person–organization fit, counsel individuals on career options, and serve as control variables in a wide variety of experiments on interventions (e.g., procedures for stress reduction). A brief outline of the major trait domains follows. An overarching distinction is made between abilities and skills (assessed with so-called maximum performance measures) and dispositions (assessed with typical performance measures).

Cognitive abilities. The value of using cognitive abilities to predict important dependent variables is well documented, and general cognitive ability (*g*) dominates (Ones, Dilchert, Viswesvaran, & Salgado, 2010; F. Schmidt & Hunter, 1998). The existence of *g* in virtually any matrix of cognitive tests and the correlation of near unity between the general factors estimated from different test batteries (e.g., see W. Johnson, Nijenhuis, & Bouchard, 2008) has been well established. The nature of the latent subfactors that make up the general factor is not a totally settled issue. The most comprehensive portrayal is still that of Carroll (1993), who acknowledged *g* as a single general factor that had eight (Carroll, 1993) or 10 (Carroll, 2003) subfactors. This portrayal is somewhat in opposition to that of Cattell (1971) and Horn (1989), who argued for the crystallized *g* and fluid *g* distinction with no general factor. Later investigations (W. Johnson & Bouchard, 2005) have tended not to support the crystallized *g*–fluid *g* structure. W. Johnson and Bouchard (2005) reanalyzed several data sets, using more sophisticated methods, and argued strongly that *g* has three subfactors: verbal, perceptual–spatial, and image rotation. However, a quantitative factor did not appear as a fourth subfactor, which might be because of the restriction of quantitative ability to simple number facility in the test batteries.

The most finely differentiated picture of how *g* could be decomposed is the comprehensive model of human abilities proposed by Fleishman and Reilly (1992), which is incorporated into O*NET (Peterson et al., 1999). It includes 21 cognitive abilities. Although some evidence has been found for differential prediction of performance across different jobs using cognitive ability subfactors (Rosse, Campbell, & Peterson, 2001; Zeidner, Johnson, & Scholarios, 1997), the incremental gains are small compared with the variance accounted for by *g*. However, even small gains are significant in the context of large-scale selection and classification in large organizations. It is also true that the advantages of using specific subfactors rather than *g* for particular measurement purposes have not been evaluated against highly specific performance subfactors (e.g., operating specific kinds of equipment that may require highly specific abilities).

Psychomotor abilities. The Fleishman and Reilly (1992) taxonomy includes 10 specific psychomotor abilities grouped into three higher order subfactors: (a) hand and finger dexterity and steadiness; (b) control, coordination, and speed of multilimb movements; and (c) complex reaction time and speed of movement involving hands, arms, legs, or all of these. Standardized performance-based tests are available for each of the 10 specific abilities, and they may (should?) be differentially important for predicting performance or specific job tasks, such as using a keyboard versus landing military jet aircraft at sea. No data are available for this domain, but it is interesting to speculate as to whether, for surgeons, open incision surgery requires somewhat different psychomotor abilities than robotic surgery.

Physical abilities. Although most occupations probably do not, several key occupations (e.g., firefighter, police officer, certain military occupations) have specialized physical ability requirements. The assessment of physical ability is also critical when considering the suitability of people with disabilities for various jobs. The latent structure of physical abilities was first investigated comprehensively by Fleishman and his colleagues (Fleishman, 1964; Fleishman & Quaintance, 1984; J. Hogan, 1991; Myers, Gebhardt, Crump, & Fleishman, 1993), who eventually arrived at a six-factor latent structure (i.e., static strength, explosive strength, dynamic strength, stamina, trunk strength, and flexibility).

Because physical ability assessment has not received as much research attention as cognitive ability assessment, at least two critical issues should be considered. First, any of the six factors may be broken down into more specific subfactors (e.g., arm and shoulder strength vs. leg strength), and for

each specific factor, there are two or more specific assessment techniques (e.g., lifting a weight off the ground vs. pushing a weight along the ground). Consequently, both the specific subfactors and the assessment method are critical choices. Gebhardt and Baker (2010) provided a thorough discussion of these issues and the research pertaining to establishing the physical requirements of work roles.

Sensory abilities. Certain occupations have specialized requirements for visual and auditory abilities (e.g., airline pilot). The Fleishman and Reilly (1992) taxonomy of sensory abilities incorporated in O*NET includes nine factors (e.g., far vision, peripheral vision, sound localization, speed recognition), each of which could be assessed by several different tests. For purposes of selection, certification, or licensure, criterion-referenced measurement is particularly critical for sensory abilities. That is, certain minimum levels of such abilities could be required, and top-down scoring would not suffice.

Somewhat strangely, the Fleishman and Reilly (1992), and consequently the O*NET, taxonomy does not include taste or olfactory abilities. Given the importance of marketing food and drink in current culture, this omission is potentially serious.

Speaking ability. O*NET includes only one such ability, speech clarity, but others may exist as well (e.g., speech modulation). Given the importance of oral communication in many occupations, this omission, too, would seem to be serious.

"Other" intelligences. The independent variable assessment landscape is also dotted with numerous variables that might be best described as "not *g*" (Lievens & Chan, 2010). The basic theme is that important abilities exist that are independent of *g* and that play a role in success at work but are not part of mainstream research. The two most prominent abilities in this category are practical intelligence (Sternberg, Wagner, Williams, & Horvath, 1995), not to be confused with a higher order construct labeled *successful intelligence* (which includes creative, analytical, and practical intelligence; Sternberg, 2003), and emotional intelligence, measured either as cognitive ability (Salovey & Mayer, 1990) or as personality (Bar-On, 1997). The available evidence pertaining to these constructs has been reviewed at some length elsewhere (Gottfredson, 2003; Landy, 2005; Lievens & Chan, 2010; Murphy, 2006). The overall conclusion must still be that construct validity is lacking for measures of these non-*g* intelligences and that they are in fact better represented by other already existing variables. For example, a recent study by Baum, Bird, and Singh (2011) evaluated a carefully constructed domain-specific situational judgment test of how best to develop businesses in the printing industry, which was then called a test of practical intelligence. With this juxtaposition, knowledge of virtually any specific domain of job-related knowledge could be labeled *practical intelligence.* What's in a name?

Traits: Dispositions

Still within the context of stable, or at least quasi-stable, traits, the I/O psychology independent variable landscape includes many constructs reflective of dispositional tendencies, that is, tendencies toward characteristic behavior in a given context. *Personality, motives, goal orientation, values, interests,* and *attitudes* are the primary labels for the different domains.

Personality. The assessment of personality dominates this landscape (Hough & Dilchert, 2010; see also Chapter 28, this volume) in terms of both the wide range of available assessment instruments (R. Hogan & Kaiser, 2010) and the sheer amount of research relating personality to a wide range of dependent variables (Hough & Ones, 2001; Ones, Dilchert, Viswesvaran, & Judge, 2007). The efficacy of personality assessment for purposes of predicting the I/O psychology dependent variables has had its ups and downs, moving from up (Ghiselli, 1966) to down (Guion & Gottier, 1965) to up (Barrick & Mount, 1991, 2005), to uncertainty (Morgeson et al., 2007), to reaffirmation (R. Hogan & Kaiser, 2010; Hough & Dilchert, 2010; Ones et al., 2007). The ups and downs are generally reflective of how the assessment of personality is represented (e.g., narrow vs. broad traits), which dependent variables are of interest, how predictive validity is estimated, and the utility ascribed to particular magnitudes of estimated validity. The bottom line is that personality assessment is a very useful enterprise so long as

the inferences that are made are consistent with the evidence pertaining to the dependent variables that can be predicted by appropriate assessments.

The assessment of personality for predictive or diagnostic purposes is complex for at least the following reasons.

- The measurement operations (i.e., "items") can come from different models of what constitutes personality description. The lexical approach is based on the words used in normal discourse to describe behavioral tendencies in others. The latent structure of such descriptors can then be investigated empirically. The five-factor model of Costa and McRae (1992) is the dominant solution. A second model would be to consult more basic theories of personality (e.g., Eysenck, 1967; Markon et al., 2005; Tellegen, 1982; Tellegen & Waller, 2000), write items reflective of the components specified by the theory, and investigate their construct validity. The advocates of the theory-based approach have argued that it produces a latent structure that is tied more closely to biological substrates (DeYoung et al., 2010). Both approaches can produce hierarchical latent structures.
- Whether the descriptors (i.e., items or scales) are obtained by data mining normal discourse or by following the specifications of a theory, assessments of an individual can be obtained via self-report or observer report. Although the bulk of personality assessment in I/O psychology is self-report, observer reports may be more predictive of various aspects of performance (e.g., Oh, Wang, & Mount, 2011). Are self-reports and observer reports different constructs? R. Hogan and Kaiser (2010) argued the affirmative and referred to self-descriptions as *self-identity* and to observer descriptions as *reputations*.
- The general agreement (DeYoung, Quilty, & Peterson, 2007) is that the lexically derived Big Five are themselves multidimensional and are composed of distinct facets. Going the other direction, combining two or three of the Big Five into higher order composite dimensions (e.g., integrity) has also been useful. DeYoung (2006) argued for two basic subfactors but rejects the existence of a general factor. Whether an assessment should use composite dimensions, factors at the Big Five level of generality, or more specific facets depends on the measurement purpose.
- At the Big Five level of generality, there is considerable agreement that the five-factor model is deficient and does not include additional important constructs such as religiosity, traditionalism or authoritarianism, and locus of control (Hough & Dilchert, 2010).

Motives or needs. Alderfer (1969), Maslow (1943), McClelland (1985), Murray (1938), White (1959), and others have offered models of the latent structure of human motives, or needs. Explicitly, or by implication, motives are defined as inner states that determine the outcomes that people strive to achieve or strive to avoid. The strength of a motive determines the strength of the striving. Different motives are associated with different classes of outcomes (e.g., outcomes that satisfy achievement needs vs. outcomes that meet social needs).

Although the distinctions between the intensity of characteristic behavioral tendencies (personality) and the strength of striving for specific outcomes (motives) are not always perfectly clear, the assessment methods have been different enough to warrant considering them separately. For example, within I/O psychology the projective techniques (ambiguous pictures) used by McClelland (1985) to assess need achievement and fear of failure and the sentence completion scales used by Miner (1977) to assess the motivation to manage are not personality scales in the sense of the NEO Personality Inventory, California Psychological Inventory, or Multidimensional Personality Questionnaire. Motive assessment has more specific referents (for more information on projective measures, see Volume 2, Chapter 10, this handbook).

Goal orientation. A very specific instantiation of motive assessment that has received increasing attention in I/O psychology is the assessment of goal orientation as it has developed from the work of Dweck and colleagues (Dweck, 1986; Elliott & Dweck, 1988). Initially, two orientations (motives) were posited in the context of training

and instruction. A performance orientation characterizes individuals who strive for a desirable final outcome (e.g., final grade). Similar to McClelland (1985), the goal is to achieve the final outcomes that the culture defines as high achievement. By contrast, a mastery or learning orientation characterizes individuals who strive to learn new things regardless of the effort involved, the frequency of mistakes, or the nature of the final evaluation. It is learning for learning's sake.

As noted by DeShon and Gillespie (2005), agreement on the nature of goal orientation's latent structure, and on whether it is a trait, quasi-trait, or state variable, is not uniform. Considerable research has focused on whether learning and performance orientations are bipolar or independent and whether one or both of them are multidimensional (DeShon & Gillespie, 2005). The answers seem to be that they are not bipolar and that performance orientation can be decomposed into performance orientation–positive—the striving toward final outcomes defined as achievement—and performance orientation–negative—the striving to avoid final outcomes defined as failure. One major implication is that performance-oriented people will avoid situations in which a positive outcome is not relatively certain and that learning-oriented individuals will relish the opportunity to try, regardless of the probability of a successful outcome. Assessment of goal orientations is still at a relatively primitive stage (Payne, Youngcourt, & Beaubien, 2007) and has not addressed the issue of whether learning or performance orientations are domain specific. For example, could an individual have a high learning orientation in one domain (e.g., software development) but not in another (e.g., cost control)? Also, the question of whether goal orientation is trait or state has not been settled. However, even though assessment is primitive, research has suggested that goal orientation is an important determinant of performance and satisfaction in training and in the work role (Payne et al., 2007).

Interests. Interest assessment receives the most attention within the individual, not the organizational, perspective and is a major consideration in vocational guidance, career planning, and individual job choice. It has also played a role, albeit smaller, in personnel selection and classification on the basis of the notion that individuals will devote more attention and effort to things that interest them, other things being equal, including the mastery of relevant skills (Van Iddekinge, Putka, & Campbell, 2011).

Assessment of interests is dominated by two inventories, the Self-Directed Search (Holland, 1994) and the Strong Interest Inventory (Harmon, Hansen, Borgen, & Hammer, 1994). The Self-Directed Search portrays interest via the now-familiar RIASEC (realistic, investigative, artistic, social, enterprising, and conventional) hexagon, which says that the latent structure of interests is composed of six factors with a particular pattern of intercorrelations. The RIASEC profiles can be used to characterize both individuals and jobs or occupations. A profile for an occupation is supposedly indicative of the degree to which the occupation will satisfy each of the six interest areas. Holland (1997) viewed the Self-Directed Search as a measure of personality and essentially subsumed interests within the overall domain of personality. The Strong Interest Inventory uses empirical weighting to differentiate individuals in an occupation from people in general on preferences for specific activities, school subjects, and so forth. Such preferences are not viewed as synonymous with personality. The Strong Interest Inventory is also scored in terms of 20 basic interest dimensions that have relatively low correlations with personality measures (Sullivan & Hansen, 2004). Whether interests account for incremental variance in the dependent variables, when compared with personality or cognitive ability, has only begun to be researched (see Van Iddekinge et al., 2011; for more information on the assessment of interests, see Volume 2, Chapter 19, this handbook).

Values. Although defining values presents the usual difficulties of choosing from among alternatives, Chan (2010) presented a careful synthesis. Values seem most usefully defined as "the individual's stable beliefs that serve as general standards by which he or she evaluates specific things, including people, behaviors, activities, and issues" (Chan, 2010, p. 321). By this specification, which distinguishes

values from personality, motives, and interests, the assessment of values can play an important role in career planning, specific job choice, and decisions to stay from the individual's perspective and in personnel selection, person–organization fit, organizational commitment, and turnover from the organization's perspective.

The latent structure of values in the context of work has not been studied very intensively. As noted by Chan (2010), the taxonomy produced by Schwartz and Bilsky (1990) is perhaps the most useful. It has 10 values dimensions for describing individuals and seven dimensions describing culture, for comparative purposes. Another structure is provided by Cooke and Rousseau (1988). In general, research on values and the development of methods for the assessment of values in the work context needs more attention in I/O psychology. Values as indicators of cultural distinctions across countries is another matter. Considerable research has been done using Hofstede's dimensions, and a comprehensive meta-analysis of these dimensions has been provided by Taras, Kirkman, and Steel (2010).

The State Side

The independent variables noted so far have been designated as trait variables that are relatively stable over the individual's work life, or at least the major portion of it. I/O psychology also deals with a complex structure of independent variables that are more statelike. That is, they are to some degree malleable, if not dynamic, as the result of situational effects, planned or unplanned. State variables are no less important than trait variables in explaining individual differences in the critical dependent variables, and the interaction between trait and state should be considered as well. The important state variables also tend to mirror the ability versus disposition distinction. That is, for some state variables, the assessment of maximum performance is the goal, whereas for others, the assessment of representative or typical dispositional states is the goal. More concretely, the distinction is between knowledge and skill versus attitudes and the cognitive regulation of choice behavior. However, for both abilities and dispositions, the distinctions between state and trait are developmentally complex. Ackerman (2000), Ackerman and Rolfhus (1999), Kanfer and Heggestad (1997), and Lubinski (2010) have provided a roadmap.

Knowledge and Skill

Specifications for knowledge and skill are elusive. What follows is an elaboration on Campbell and Kuncel (2001) and an attempt to distinguish among (a) declarative knowledge, (b) proceduralized knowledge, (c) skill, and (d) problem solving. It is meant to be consistent with Anderson (1987) and Simon (1992). The nature of competencies is a separate issue.

Declarative knowledge is knowledge of labels and facts pertaining to objects, events, processes, conditions, relationships, rules, if–then relationships, and so forth. As in the Anderson (1987) framework, declarative knowledge is distinguished from *proceduralized knowledge,* which refers to knowing how something should be done (e.g., How should shingles be put on a roof? How should a correlation matrix be factor analyzed? How should a golf club be swung?). In contrast to knowing how to do something, *skill* refers to actually being able to do it. Sometimes the distinction between proceduralized knowledge and skill is relatively small (e.g., knowing how to factor analyze a matrix vs. actually doing it), and sometimes it is huge (e.g., knowing how to swing a three-iron and actually being able to do it at some reasonable level of proficiency; note the qualifier—skills are not dichotomous variables). Consequently, a skill can be defined as the application of declarative and proceduralized knowledge capabilities to solve structured problems and accomplish specified goals. That is, the problems or goal accomplishments at issue have known (i.e., correct) solutions and known ways of achieving them. The issue is not whether the problems or specified goals are easy or difficult, it is whether correct solutions can be specified.

The capabilities commonly labeled as *problem solving, critical thinking,* or *creativity* should be set apart from a discussion of knowledge and skill. Although these capabilities appear frequently in competency models and other forms of knowledge, skills, and abilities lists, they are seldom, if ever, given a concrete specification, seemingly because

everyone already knows what they are. Consequently, whether problem solving, creativity, and critical thinking are intended as trait or state variables is not clear. That is, are they distinct from general cognitive ability, and can they be enhanced via training and experience? Attempts to assess these capabilities must somehow deal with this lack of specification.

Following Simon (1992), *problem solving* could be defined as the application of knowledge and skill capabilities to the development of solutions for ill-structured problems. *Ill-structured problems* are characterized as problems for which the methods and procedures required to solve them cannot be specified with certainty and for which no correct solution can be specified a priori. Generating solutions for such problems is nonetheless fundamentally and critically important (e.g., What should be the organization's research and development strategy? What is the optimal use of training resources? How can the coordination among teams be maximized?). Specified in this way, a problem-solving capability is important for virtually all occupations, which invites a discussion of how it can be developed and assessed. The literature on problem solving within cognitive psychology in general, and with regard to the study of expertise in particular, is reasonably large (Ericsson, Charness, Feltovich, & Hoffman, 2006). To make a long story brief, the conclusions seem to be that (a) there is no general (i.e., domain-free) capability called problem solving that can be assessed independently of *g;* (b) problem-solving expertise, as defined earlier, is domain specific; (c) expert problem solvers in a particular substantive or technical specialty simply know a lot, and what they know is organized in a framework that makes it both useful and accessible; and (d) experts use a variety of heuristics and cues correctly to identify and structure problems, determine what knowledge and skills should be applied to them, and judge which solutions are useful.

Currently, expert problem solving is viewed as a dual process (Evans, 2008). That is, solutions are either retrieved from memory very quickly, seemingly with minimal effort and thought, or a much more labor-intensive process of problem exploration and definition occurs, thinking about and evaluating potential solutions and finally settling on a solution or course of action. The latter process is not a serial progression through a specific series of steps, but it is an organized effort to use the expert's fund of knowledge, skills, and strategies in a useful way.

The dual-process models are not strictly analogous to automatic versus controlled processing distinctions (Ackerman, 1987). The distinction is more between identifying a solution very quickly versus identifying one more deliberately. Different brain processes are involved, as evidenced by functional magnetic resonance imaging studies (Evans, 2008). Some investigators (e.g., Salas, Rosen, & DiazGranados, 2010) have been quick to label the fast process *intuition* and insert it into competency models, knowledge, skills, and abilities lists, and the like—again with virtually no specifications for what intuition is. It is another example of an important word from general discourse causing assessment problems for I/O psychology when attempting to incorporate it in research or practice.

Following Simon (1992), Kahneman and Klein (2009) demystified intuition by defining it as a process that occurs when an ill-structured problem to be solved exists, and the problem situation provides cues that the expert can use to quickly access relevant information stored in memory that provides a useful solution. Virtually by definition, intuitive expertise must be based on a large, optimally structured base of information and on identifying the most valid situational cues. There is no magic in intuition. With regard to solving ill-structured problems, the distinction between quickly accessing a useful solution (i.e., intuition) and being more deliberative is not a clear dichotomy. A final solution might be produced quickly but then subjected to varying degrees of deliberation.

Solving structured problems (i.e., exhibiting a skill as defined earlier) is a somewhat different phenomenon. Certain (but certainly not all) skills can be practiced enough so that they do become automatic (Ackerman, 1988) and can be used without effort or conscious awareness. However, many skills will always remain a controlled or deliberative process (e.g., creating syntax). Experts do it more quickly and more accurately than other people, but not automatically.

Creativity

What then are creativity and critical thinking? Answering such questions in detail is beyond the scope of this chapter, but the following discussion seems relevant vis-á-vis their assessment. Comprehensive reviews of creativity theory and research are provided by Dilchert (2008), Runco (2004), and Zhou and Shalley (2003).

Creativity has been assessed as both a cognitive and a dispositional trait, as in creative ability and creative personality. Both cognitive- and personality-based measures have been developed via both empirical keying (e.g., against creative vs. noncreative criterion groups) and homogeneous, or construct-based, keying. Meta-analytic estimates of the relationships between cognitive abilities and creative ability and between established personality dimensions (e.g., the Big Five) and creative personality scales are provided by Dilchert (2008) as well as the correlations of creative abilities and creative personality dimensions with measures of performance.

Within a state, framework creativity can also be viewed as a facet of ill-structured problem-solving performance (e.g., George, 2007; Mumford, Baughman, Supinski, Costanza, & Threlfall, 1996). Here, the difficulty is in distinguishing creative from noncreative solutions. The specifications for the distinction tend not to go beyond stipulating that creative solutions must be both unique, or novel, and useful (George, 2007; Unsworth, 2001). That is, uniqueness by itself may be of no use. In the context of problem-solving performance, is a unique (i.e., creative) solution just another name for a new solution, or is it a distinction between a good solution and a really good solution (i.e., the latter has more value than the former, given the goals being pursued)? In general, creativity as a facet of a problem-solving capability does not seem unique. Attempting to assess creative expertise as distinct from high-level expertise may not be a path well chosen.

Critical Thinking

Similar specification problems characterize the assessment of critical thinking, which has assumed rock-star construct status in education, training, and competency modeling (e.g., Galagan, 2010; Paul & Elder, 2006; Secretary's Commission on Achieving Necessary Skills, 1999; Stice, 1987). Many, many definitions of critical thinking have been offered in a wide variety of contexts ranging from the Socratic tradition, to the constructivist perspective in education, to economic theory, to problem solving in the work role, to the value-added assessment of education, and to the scientific method itself. In all of these, critical thinking is regarded, explicitly or implicitly, as a state variable. That is, it is something to be learned. Moreover, it could be regarded as a cognitive capability or as a motivational disposition (i.e., people differ in the degree to which they want to think critically). Perhaps the former is a prerequisite for the latter.

Setting aside those specifications that are so general as to be indistinguishable from thinking, problem solving, or intelligence itself, the defining characteristic of critical thinking seems to be a disposition to question the validity of any assertion about facts, events, ongoing processes, forecasts of the future, and so forth and to ask why the assertion was made. The form of the questioning (i.e., critical thinking) relies on the canons of rationality, logic, and the scientific method and on domain-specific knowledge. That is, to think critically is to always question the truth value of a statement (a disposition) and to analyze (a cognitive capability) the basis on which the statement is made.

Such a specification invites a consideration of whether such a thing as a general critical thinking skill exists, or whether it must always be substantially domain specific. That is, is it even possible to talk about critical thinking independently of content domain? This is the same issue discussed earlier in the context of problem-solving capabilities and creativity.

The assessment of critical thinking is most often via rater judgment and less often by standardized tests (Ennis, 1985; Ewell, 1991; Steedle, Kugelmass, & Nemeth, 2010). One area of research that has confronted both the general versus domain-specific issue and rated versus tested assessment is the development of the value-added approach to the assessment of educational outcomes (Liu, 2011). This effort has been in progress for some 30 or more years but has surged recently as a means for assessing teacher effects (kindergarten–Grade 12) on

student achievement and the college–university effect on undergraduate learning (Klein, Freedman, Shavelson, & Bolus, 2008). The latter is perhaps more relevant and involves the assessment of gains on certain general skills—critical thinking being a major one—as a function of a college or university education. Three principal assessment systems are available (Banta, 2008): the Collegiate Assessment of Academic Proficiency from American College Testing, the Measure of Academic Proficiency and Progress from the Educational Testing Service, and the Collegiate Learning Assessment from the Council for Aid to Education. The first two have a multiple-choice format, but the third uses open-ended (i.e., written) responses to three scenarios involving (a) taking and justifying a particular position on an issue, (b) critiquing and evaluating a particular position on an issue, and (c) performing the tasks in an in-basket simulation. The responses are scored by expert raters to yield scores on problem solving, analytic reasoning, critical thinking, and writing skills. The stated expectation is that the college or university experience should increase such skills, and schools can be ranked in terms of the extent to which they do so (Klein et al., 2008). Research so far has suggested that scores on such measures do go up from freshman to senior status, but it has been difficult to extract more than one general factor, and the construct validity of the general factor has not been clearly established.

The moral here is that for assessment purposes, problem solving, creativity, and critical thinking are complex and extremely difficult constructs to specify. They are particularly difficult to specify in a domain-free context. Moreover, is the domain-free context even the most relevant for assessment in I/O psychology? These issues should not be approached in a cavalier fashion, such as listing them in a competency model without thorough specification.

Latent Structure of Knowledge and Skills (as Determinants of Performance)

For the assessment of individual differences in domain-specific knowledge and skills, a distinction can be made between the direct real-time knowledge and skills determinants of performance in a work role and the knowledge and skills requirements that are assessed before being hired. The former might be assessed for diagnostic or developmental purposes and the latter for predictive purposes. However, the latter may also serve as a prerequisite for the former and, as asserted in a previous section, the latter (indirect) can only influence performance by influencing the former (direct).

In contrast to abilities, the substantive latent structure or structures of knowledge and skills have received scant attention. Part of the problem is simply the almost limitless number of possibilities and the difficulty of choosing the appropriate levels of generality or specificity. That is, many, many knowledge and skills domains exist, and they may be sliced very coarsely or very finely.

Content-based knowledge taxonomies do exist. A relatively general one is included in O*NET and consists of 38 knowledge domains that are primarily focused on undergraduate curriculum areas (e.g., psychology, mathematics, philosophy, physics). As noted by Tippins and Hilton (2010), the knowledge requirements for many skilled trades, or technical specialties not requiring a college degree, do not seem to be represented. A taxonomy-like structure that does represent the non–bachelor's degree specialties is the compilation of technical school curricula known as the Catalog of Instructional Programs maintained by the U.S. Department of Education.

Knowledge taxonomies specific to particular classes of occupations have also been developed via comprehensive job analysis efforts over a period of years by the U.S. Office of Personnel Management (2007). To date, they cover these classes of occupations:

- professional and administrative,
- clerical,
- technical,
- executive or leadership,
- information technology, and
- science and engineering.

Collectively, they are a part of the Office of Personnel Management's MOSAIC system and constitute a much more complete taxonomy of job knowledge requirements than the O*NET.

Portraying the taxonomic structure for direct and indirect skills requirements is even more

problematic than it is for knowledge. O*NET provides a taxonomy of 35 skills that are defined as cross-occupational (i.e., not occupation specific) and that vary from the basic skills such as reading, writing, speaking, and mathematics, to interpersonal skills such as social perceptiveness, and to technical skills such as equipment selection and programming. As noted by Tippins and Hilton (2010), the O*NET skills are very general in nature and generally lacking in specifications. Moreover, two of the 35 O*NET skills are complex problem solving and critical thinking, the limitations of which were discussed earlier. Again, a wider set of more concretely specified skills are included in the Office of Personnel Management's MOSAIC system but only for certain designated occupational groups.

Because the skills gap has been such a dominant topic in labor market analyses (e.g., Davenport, 2006; Galagan, 2010; Liberman, 2011), one might expect the skills gap literature to provide an array of substantive skills that are particularly critical for assessment. It generally does not. Virtually all skills gap information is obtained via employer surveys in response to items such as "To what extent are you experiencing a shortage of individuals with appropriate technical skills?" However, the specific technical skills in question are seldom, if ever, specified. Skills such as leadership, management, customer service, sales, information technology, and project management are as specific as it seems to get.

The purposes for which knowledge and skill assessments might be done are, of course, varied. It could be for selection, promotion, establishing needs for training and development, or certification and licensure—all from the organizational perspective. From the individual perspective, it could be for purposes such as job search, career guidance, or self-managed training and education. For organizational purposes, the lack of a taxonomic structure may not be a serious impediment. Organizations can develop their own specific measures to meet their needs, such as specific certification or licensure examinations. However, for individual job search or career planning purposes, the lack of a concrete and substantive taxonomic structure for skills presents problems. Without one, how do individuals navigate the skills domain when planning their own education and training or matching themselves with job opportunities?

State Dispositions

By definition, and in contrast to trait dispositions, state dispositions are a class of independent variables that determines volitional choice behavior in a work setting but that can be changed as a result of changes in the individual's environment. Disposition-altering changes could be planned (e.g., training) or unplanned (e.g., peer feedback). A selected menu of such state dispositions follows.

Job Attitudes

There are many definitions of *attitudes* (Eagley & Chaiken, 1993), but one that seems inclusive stipulates that attitudes have three components: First, attitudes are centered on an object (e.g., Democrats, professional sports teams, the work you do); second, an attitude incorporates certain beliefs about the object (e.g., Democrats tax and spend, professional sports teams are interesting, the work you do is challenging); and third, on the basis of one's beliefs, one has an evaluative–affective response to the object (e.g., Democrats are no good, professional sports teams are worth subsidizing, you love the challenges in your job). The evaluation–affective reaction is what influences choice behavior (e.g., you vote Republican, you vote for tax subsidies for a professional sports stadium, you will work hard on your job for as long as you can).

Job satisfaction. The job attitude that has dominated both the I/O research literature and human resources practice is of course job satisfaction, which was discussed earlier in this chapter as a dependent variable. However, used as an independent variable the correlation between job satisfaction and both performance and retention has been estimated literally hundreds of times (Hulin & Judge, 2003) using the same assessment procedures discussed previously, and the same issues apply (e.g., Weiss, 2002). In addition to job satisfaction, several other work attitudes have received attention for both research and application purposes.

Commitment. As an attitude, commitment in a work setting can take on any one of several different

objects, and it is possible to assess commitment to the organization, the immediate work group, an occupation or profession, one's family or significant other, and entities outside of the work situation such as an avocation or civic responsibility. Beliefs about any one of these attitude objects could lead to positive or negative affect that influences decisions to commit effort for short- or long-term durations. The assessment issues revolve around the differentiation of attitude intensity across objects and the distinction between *commitment to* and *satisfaction with.* That is, measures of job satisfaction and organizational commitment both yield significant correlations with turnover and performance (Hulin & Judge, 2003), but does one add incremental variance over the other (Crede, 2006)? Both the latent structure of commitment and its distinctiveness from other attitudes are not settled issues.

Job involvement. Job involvement is variously characterized as a cognitive belief about the importance of one's work, the degree to which it satisfies individual needs of a certain kind (e.g., achievement, belongingness), or the degree to which an individual's self-identity is synonymous with the work he or she does (Brown, 1996; Kanungo, 1982; Lodahl & Kejner, 1965). Consequently, it should be related to job satisfaction, self-assessments of long-term performance, commitment to the occupation (but perhaps not the organization), and intentions to stay or leave.

Job engagement. Job engagement is currently a hot topic, as evidenced by at least two recent handbooks (Albrecht, 2010; Bakker & Leiter, 2010) and a major book (Macey, Schneider, Barbera, & Young, 2009). In their focal article in the journal *Industrial and Organizational Psychology: Perspectives on Science and Practice*, Macey and Schneider (2008) made a concerted attempt to define *state engagement,* which was characterized as an evaluative or affective state regarding one's job that goes beyond simply being satisfied, committed, or involved and reflects the individual's total passion and dedication for his or her work and a willingness to be totally immersed in it. The article elicited 13 quite varied responses that illustrated the major assessment issues with which such constructs must deal, in both research and practice. For example, what is the latent structure of this construct? Is engagement a dispositional trait, and affective state, or a facet of performance itself? Do measures of engagement account for unique variance over and above satisfaction and commitment? Although managements tend to view engagement as an important construct (Masson, Royal, Agnew, & Fine, 2008; Vosburgh, 2008), its assessment must deal with the preceding issues. Christian, Garza, and Slaughter (2011) reported a meta-analysis that engages some of the issues. Although the number of studies is not great, and there is variation in the measures of engagement, the evidence is supportive of unique variance and some incremental predictive validity that could be attributed to engagement (for further discussion of job satisfaction and related job attitudes, see Chapter 37, this volume).

Motivational States

Again, in contrast to trait dispositions, such as the need for achievement, a class of more dynamic motivational states has become increasingly important, at least in the research literature, as determinants of choice behavior at work. Consider the following sections.

Self-efficacy and expectancy. The Bandurian notion of self-efficacy is the dominant construct here and is defined as an individual's self-judgment about his or her relative capability for effective task performance or goal accomplishment (Bandura, 1982). Self-efficacy judgments are specific to particular domains (e.g., statistical analysis, golf) and can change with experience or learning. Self-efficacy is similar to, but not the same as, Vroom's (1964) definition of *expectancy* as it functions in his valence–instrumentality–expectancy model of motivated choice behavior. *Expectancy* is an individual's personal probability estimate that a particular level of effort will result in achieving a specific performance goal. It is very much intended as a within-person explanation for why individuals make the choices they do across time, even though it is most frequently used, mistakenly, as a between-persons assessment.

Instrumentality (risk) and valence (outcome value). From subjective expected utility to valence–instrumentality–expectancy theory (Vroom,

1964) to prospect theory (Kahneman & Tversky, 1979), the concepts of risk assessment and outcome value estimation are viewed as state determinants of choice behavior. Individuals want to minimize risk and maximize outcome value and will govern their actions accordingly. However, as noted in prospect theory, preference for risk levels and outcome values are discounted as a function of time. That is, individuals will take on greater risk but value specific outcomes less the farther they are in the future. See Steel and Konig (2006) for an integrated summary of how such state dispositions influence choice behavior. Such considerations have not yet played a very large role in diagnostic assessment of the choice to perform, but perhaps they should.

Core self-evaluation. Judge, Locke, Durham, and Kluger (1998) have done considerable work on a set of dispositions they referred to as *core self-evaluations*. The set consists of general self-efficacy (i.e., a self-assessment of competence virtually regardless of the domain), self-esteem, locus of control, and neuroticism. It is somewhat problematic as to whether these facets can be considered trait or state, but they have shown significant predictive validities (Judge, Van Vianen, & DePater, 2004). Their distinction as separate facets is also not a settled issue and may depend on the specific measure involved (Ferris et al., 2011; Judge & Bono, 2001).

Mood and emotion. The dispositional effects of mood and emotion on work behavior have received increasing attention (Mitchell & Daniels, 2003; A. M. Schmidt, Dolis, & Tolli, 2009; Weiss & Rupp, 2011). Specifications for these constructs are not perfectly clear (Mitchell & Daniels, 2003), but in general, mood is defined as an affective state that is quite general, and emotion is usually specified as having a specific referent. That is, one's mood is generally bad or good, but the individual is emotional (positively or negatively) about specific things. Why assess such dispositional states? The dominant answer is that as state determinants of choice behavior, they help to explain the within-person variability in performance over relatively short periods of time (e.g., Beal et al., 2005). Also, as advocated by Weiss and Rupp (2011), the whole person cannot be assessed without a consideration of these states.

Things that are known to be unknown. A list of state determinants is probably not complete without noting that individuals are not aware of all of the determinants of their choice behavior (e.g., Bargh & Chartrand, 1999). That is, people make many choices, even at work, for which they cannot explain the antecedents. Apparently, the reasons for action are not in conscious awareness. Can they be recovered via some form of assessment? That has yet to be determined, but one avenue of investigation concerns priming effects (Gollwitzer, Sheeran, Trotschel, & Webb, 2011).

Competencies (and Competency Modeling)

So far, this chapter has avoided the question of whether competencies and competency modeling are, or are not, a distinct sector of the I/O psychology assessment landscape. That is, is competency modeling just knowledge, skills, abilities, and other characteristics (KSAOs) and job analysis by another name, or should it be set apart? Previous attempts to settle this question have been inconclusive (e.g., Sackett & Laczo, 2003; Schippman, 2010; Schippman et al., 2000). In a further attempt at clarity, Campion et al. (2011) outlined best practices in competency modeling and noted its most distinctive features, in the context of the following definition of competencies. That is, *competencies* are defined as individual KSAOs, or collections of KSAOs, that are needed for effective performance of the job in question. By this definition, competencies are determinants of performance, not performance itself. Unfortunately, Campion et al.'s most detailed example of a competency (p. 240) is of project management, the specifications for which seem to be a clear characterization of performance itself, such that the example is not consistent with the definition. The competency modeling literature has variously referred to knowledge, skills, abilities, personal qualities, performance capabilities, and many other things (e.g., attitudes, personality, motives) as competencies (Parry, 1996). In the aggregate, very little of the I/O psychology landscape is left out, and Clouseau's dictum potentially complicates assessment—that is, if competencies are everything, then they risk being nothing.

Campion et al. (2011) attempted to keep that from happening by abstracting best practices and identifying what makes competency modeling unique. Perhaps their most salient points are the following:

- Ideally, competency models attempt to develop specifications for the levels of a competency that distinguish high performers from average or low performers. That is, to paraphrase, how do the performance capabilities of expert performers differ substantively from the performance capabilities of nonexpert performers, and what level of knowledge, skills, and dispositional characteristics are required to exhibit expert performance levels? This is very different from conventional job analysis, which tries to identify the components (e.g., tasks, work activities) of performance and predict which KSAOs will be correlated (or linked) with them. However, competency modeling is similar to cognitive job analysis (Schraagen, Chipman, & Shalin, 2000), which asks how experts, when compared with novices, perform their jobs and what resources (e.g., knowledge, skills, strategies) do they use to perform at that level? Cognitive job analysts and competency modelers should interact more. They have things in common.
- High-level subject matter experts (e.g., executives) are used to first specify the substantive goals of the enterprise and then identify (to the best of their ability) both the performance and KSAO competencies at each organizational level that will best facilitate goal accomplishment. This is in contrast to conventional job analysis, which asks incumbents or analysts to rate the importance of KSAOs for performance in a target job, without reference to the enterprise's goals. Supposedly, the incumbent or analyst subject matter experts (SMEs) have these in mind when making linkage judgments, but perhaps not.
- If competencies are specified as in the first bullet, the various components of the human resources system can more directly address enterprise objectives by focusing selection, training, and development on obtaining the most critical competencies. In some respects, competency modeling is analogous to a needs analysis.

Even from this brief examination, it is apparent that competency modeling carries a heavy assessment burden. This burden is complicated by a resistance to taxonomic thinking and a desire to specify competencies in organizational language. These choices may aid in selling competency modeling to higher management, but they complicate specification for assessment. For example, how can previous theory and research in leadership be used to define and specify performance levels for leading with courage? Such a competency has a nice ring to it, but what does it mean? Tett, Guterman, Bleier, and Murphy (2000) attempted to address some of these issues by beginning with the research literature and conducting a systematic content analysis of published management competencies intended to reflect performance capabilities. On the basis of SME judgments, they identified 53 competencies grouped into 10 categories and attempted a definition of each of them. Although this effort represents a significant step in the right direction, a few of the competencies still seem more like personality characteristics than performance capabilities (e.g., orderliness, tolerance). However, their juxtaposition of the SME-developed taxonomy derived from the literature against the competency lists from several private firms is interesting.

THE CONTEXT

So far, this chapter, in the interests of demonstrating the complexity of assessment in I/O psychology, has tried to outline the basic elements in the dependent and independent variable landscape that invite measurement. Because the concern is assessment, the complexities of research and practice focused on estimating the interrelationships among, or between, independent and dependent variables, differential prediction across criteria and interactive effects are not addressed. These are questions that, although very critical, do not themselves change the measurement requirements for the variables involved.

However, I/O psychology does make a big deal of the influence of the context, or situation, on the interrelationships among variables. Such contextual variables are often referred to as *moderators*. Also,

the context can take on the status of an independent variable. For example, the organizational climate or culture might be hypothesized to influence individual choice behavior. Consequently, it is sometimes important to assess the context itself. For example, Scott and Pearlman (2010) strongly made the case that assessment for organizational change must always deal with assessment of the context.

The literature on the assessment of the context is in fact very large. For example, in the course of developing the specifications for the O*NET database two taxonomies were created, one for the work (job) context (Strong, Jeanneret, McPhail, Blakley, & D'Egidio, 1999) and one for the organizational context (Arad, Hanson, & Schneider, 1999). They are both multilevel hierarchical taxonomies. The work context is portrayed as having 39 first-order factors and 10 second-order factors, such as how people in work roles communicate, the position's environmental conditions, the criticality of the work role, and the pace of the work. The organizational context is reflected by 41 first-order dimensions and seven second-order factors such as organizational structure, organizational culture, and goals.

Not surprisingly, because they were based on extensive literature searches, the O*NET's work and organizational context taxonomies subsume much of the literature on organizational culture and climate (e.g., James & Jones, 1974; Ostroff, Kinicki, & Tamkins, 2003), organization development (J. R. Austin & Bartunek, 2003), and work design (J. R. Edwards, Scully, & Bartek, 1999; Morgeson & Campion, 2003). Within O*NET, the context is assessed via job incumbent ratings. Although a detailed examination of the context literature cannot be presented here, the major features of the context that dominate the need for assessment, and the issues that assessment of the context creates, seem in the author's opinion to be as follows.

1. the features of the work context that are identified as rewarding or need fulfilling, such as the 20 potential reinforcers assessed by the instrumentation of the Minnesota theory of work and adjustment or the five job characteristics specified by Hackman and Oldham's (1976) Job Diagnostic survey;
2. the full range of performance feedback provided by the job and organizational context;
3. the nature and quality of the components of the organization's human resources system such as selection procedures, compensation practices, and training opportunities;
4. the nature of the organization's operating goals, such as those resulting from an application of Pritchard's productivity measurement system (Pritchard, Holling, Lammers, & Clark, 2002);
5. leadership emphasis, in terms of whether it is directive versus participative, formalized versus informal, or centralized versus decentralized;
6. the complexity and variety of the technologies used by the organization;
7. the relative criticality or importance of specific jobs, positions, or roles;
8. the level of conflict among work roles or units;
9. the relative pace of work in terms of the characteristic levels of effort, intensity, and influences of deadlines;
10. the physical nature of the environment (e.g., temperature, illumination, toxicity); and
11. the organizational climate and culture.

Number 11 perhaps deserves special mention. The assessment of organizational climate and culture are important topics in I/O psychology and have a long history (James & Jones; 1974; Lewin, 1951; Litwin & Stringer, 1968; Trice & Beyer, 1993). However, developing clear specifications for what constitute organizational culture and climate has proven elusive (Denison, 1996; Ostroff et al., 2003; Verbeke, Volgering, & Hessels, 1998). Verbeke et al. (1998) surveyed the published literature and identified 32 distinct definitions of climate and 54 definitions of culture. However, a not-uncommon distinction is as follows.

Organizational culture refers to the informal rules, expectations, and norms that govern behavior, in addition to written policies, that are both relatively stable and widely perceived. *Organizational climate* generally refers to individual perceptions of the impact of the work environment on individual well-being (e.g., see James & Jones, 1974). By convention, *psychological climate* refers to each individual's judgment, whereas *organizational climate* refers

to the aggregate (e.g., mean) judgment across individuals.

Besides the definitional problems, the assessment of culture and climate must deal with at least the following issues as well. First, to what unit are culture or climate referenced? Is it work group, department, division, or organizational climate or culture? Second, are individuals asked to provide their own individual judgments about the nature of the climate or culture or to predict the judgments of other organizational members? With either method, the construct of culture and climate requires some degree of consensus or agreement among individuals, but how much? Finally, is there a genuine latent structure of distinctive subfactors for culture and climate, or should both climate and culture be tied to any number of specific referents that would not necessarily constitute a taxonomy of latent dimensions? James and Jones (1974) argued for the former, and Schneider (1990) argued for the latter. Standardized survey questionnaires do exist for culture (e.g., Cooke & Rousseau, 1988) and for climate (Ostroff, 1993), and they tend to yield stable factor structures. Some evidence also exists for a general climate factor that seems to represent the overall psychological safety and meaningfulness of the work environment (Brown & Leigh, 1996). The bottom line is that any attempt to assess organizational culture and climate, either as moderator variables or as independent variables in their own right, must address these issues. As always, settling specification and assessment issues must come before considering what mediates the relationship between culture or climate and something else, or the boxes, arrows, and path coefficients have little meaning.

Psychometric Landscape

Many features of the psychometric landscape, as they pertain to measurement and assessment in I/O psychology, are well known and have not been discussed, yet again, in this chapter. For several assessment purposes, psychologists are governed by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999) and the Society of Industrial and Organizational Psychology's (2003) *Principles,* and all professionals are familiar with them. Also, all appropriate professionals should be familiar with the development of measurement theory beyond the confines of Spearman's (1904) classic model of true and error scores, which becomes a special case of the generalizability model (e.g., Putka & Sackett, 2010), and with the basics of item response theory (IRT) as well (Embretson & Reise, 2000).

The most important principle from the psychometric landscape is that all assessment, whether for research, practice, or high-stakes decision making, must have evidence-based validity for the purpose or purposes for which it is to be used. This principle is as true for asserting that self-efficacy is being measured in a research study or for stating that critical thinking is a required competency in a competency model as it is for using a personality measure in high-stakes personnel selection. A large literature is also available on what kinds of evidence support the various purposes for which assessment is done (e.g., see AERA et al., 1999; Farr & Tippins, 2010; McPhail, 2007; Scott & Reynolds, 2010). This literature should be part of all I/O psychologists' expert knowledge base and is discussed in Chapter 4 of this volume. However, for a somewhat contrarian view, see Borsboom, Mellenbergh, and van Heerden (2003) and Borsboom (2006).

There are also some less talked-about issues that readers should think about. The first challenges the very existence of applied psychology. It comes primarily from the work of Michell (1999, 2000, 2008) and others (e.g., Kline, 1997) who asserted that psychometrics (i.e., measurement in psychology) is a pathological science. For them, measurement in psychology is pathological for two reasons. First, virtually all constructs studied or used in psychology are not quantitative but are simply assumed to be so without further justification. In this context, being quantitative essentially means that the scores representing individual differences on a variable constitute at least an interval scale. Second, the lack of justification for the assumption of such scale properties is kept hidden (i.e., never mentioned). Consequently, what can be inferred about individual differences on nonquantitative variables, and what

do the relationships (e.g., correlations) among such variables actually mean? For example, if psychologists assess training effects by administering an achievement test before and after training and report that training produced a gain of 0.5 standard deviations (i.e., $d = 0.50$), what does that mean in terms of what or how much was learned? If neither job satisfaction nor job performance are assessed on a scale with at least interval properties, what does an intercorrelation of .35 mean?

This issue is an old one and goes back to the bifurcation between Stevens (1946), who asserted that measurement is the assignment of numbers to individuals according to rules, and Luce and Tukey (1964), who counterargued for a conjoint measurement model that requires interval scales with additive properties. The current version of the argument is discussed in a series of articles by Michell (2000, 2008), Borsboom and Mellenbergh (2004), and Embretson (2006).

Everyone would probably admit that psychologists seldom deal with interval property measurement, and the response to the accusation of pathology could be one of four kinds. First, it might be argued that ordinal scales are okay for many important assessment purposes (e.g., top–down selection). Second, the purpose of assessment may not be to scale individuals but to provide developmental feedback. Third, many of the variables psychologists study are quantitative, because when the same variable is measured with different instruments, the results are the same. The assumption has just not been explicitly tested. Fourth, one could argue that psychologists do, on occasion, assess people quantitatively, as in criterion-referenced measurement (Cizek, 2001) or when using IRT models (Embretson, 2006). Borsboom and Mellenbergh (2004) argued that the Rasch model (i.e., a one-parameter IRT model) is an essentially stochastic equivalent to the deterministic conjoint measurement model, because it simultaneously scales both items and individuals on the same scale (i.e., theta).

The preceding issue is related to the recent discussion of dominance versus ideal-point scaling for attitude and personality assessment (Drasgow, Chernyshenko, & Stark, 2010). In psychology, these two scaling procedures are credited to Likert (1932) and Thurstone (1928), respectively. Thurstone scaling does provide information about the relative size of the intervals between scores on the attitude–personality continuum. Drasgow et al. (2010) argued persuasively that embedding ideal-point scaling in an IRT model overcomes some of the previous difficulties in Thurstone's scaling and results in a more quantitatively scaled variable. This application has also been used for performance assessment (Borman et al., 2001) via computer-adaptive rating scales.

Another measurement-related criticism of assessment in I/O psychology is that the field has seemed to show little interest in test taking as a cognitive process. That is, I/O psychologists do not ask questions about how a test taker decides on a particular response and cannot give a cognitive account of the processes involved (e.g., Mislevy, 2008). The implication is that two individuals may arrive at the same response in different ways (Mislevy & Verhelst, 1990), which in turn implies that their scores do not mean the same thing. This criticism is most often made in the context of ability or achievement testing, but it could also be directed at attitude measurement, linking judgments in job analysis, assessor ratings in assessment centers, and performance ratings in general.

A final issue, and perhaps the most important one, concerns how the structure of the various domains of dependent, independent, and situational variables should be modeled. A very thorough and sophisticated treatment of latent and observed structures was provided by Borsboom et al. (2003). They discussed three distinct ways to model the covariance structure of a set of observed scores as a function of latent variables. In the first model, latent variables are constructs that cause responses to operational measures but are not equivalent to them. For example, general mental ability is a latent variable that most surely has neurological substrates, as yet unknown, that were formed by heredity, experience, and their interaction. The existence of the latent variable is inferred from the covariances of the measures constructed to measure it. The observed covariances using a variety of such measures always yield a general factor. Corrected for attenuation, the intercorrelations of scores on the general factor when obtained from independent sets

of tests approach unity. This example is the clearest of a real latent variable. Also, it does not preclude the existence of subfactors (e.g., verbal, quantitative) that yield highly predictable covariances among observed scores. The latent structure of other trait domains is not quite as clear, at least not yet, but the evidence is sufficient to suggest that such a latent structure exists for some of them, such as personality and interests. In fact, much of the work on the latent structure of personality is an attempt to map the biological substrates of the factors (DeYoung, 2006).

A second model, at the other extreme, is to assert that observed factor scores are nothing more than the sum of the individual scores (i.e., items, tests, ratings) that compose them. Borsboom et al.'s (2003) example is from sociology. Suppose, for an individual, socioeconomic status is taken as the sum of income level, education level, and home value. There is no latent variable labeled *socioeconomic status* that determines income, education, and home value. It can only be defined in terms of the three operational measures. That is not to say that the sum score labeled socioeconomic status is not valuable; however, it does represent a different model that cannot be used in the same way as a substantive latent trait model. Consequently, every time socioeconomic status is used as a label for a sum score, the specific measures being aggregated must be spelled out. If there are correlations among the specific measures, they must be explained by common determinants from other domains (e.g., general mental ability and conscientiousness).

The third model represents the attack of the postmodernists on the generally realist approach to research and practice that characterizes applied psychology (Boisot & McKelvey, 2010; P. Johnson & Cassell, 2001). That is, observed covariance structures are social constructions that result from how researchers or practitioners construct the way they observe organizational behavior. The postmodernists have asserted that assessment in research and practice cannot be independent of this personal psychology. Agreement on such social constructions results from the socialization and training processes in I/O psychology. There really is no such thing as an independent latent variable (construct) that determines the covariance structure of observed measures.

Models 2 and 3 are more similar to each other than they are to the first model, and for I/O psychology the basic issue is when should Model 1 versus Model 2 be invoked. Depending on the choice, the structural equations are different, and the analysis procedures are different (MacKenzie, Podsakoff, & Jarvis, 2005; Podsakoff, MacKenzie, Podsakoff, & Lee, 2003). Some additional implications of model choice are at least those described next.

If it is appropriate to model the trait determinants of performance (e.g., cognitive ability, personality, motives) as a function of latent variables, then it is appropriate, and necessary, to base assessment on the specifications for the latent variables. Constantly inventing new variables without reference to a known or specified latent structure is dysfunctional for research and practice.

In contrast to trait assessment, imposing a latent variable model on state assessment is more problematic. For example, are there skill and knowledge domains that can be specified well enough that testing and assessment can estimate a domain score that has surplus meaning beyond the sum of a particular set of item scores? This is one thing that made development of knowledge and skill taxonomies for O*NET difficult. However, IRT models provide a way of testing whether latent models are reasonable. A similar question could be asked about attitude or climate assessment. For example, are there general (latent?) dimensions of organizational climate, or should climate always be referenced to specific organizational activities or procedures? Also, what does a path analysis actually estimate if a latent variable model is not appropriate?

These considerations raise another obvious question. That is, what is the latent structure of performance itself, or is there one, and what is the impact of this issue on assessment? In this regard, some things are certain, some things are reasonably certain, and some are currently indeterminate. For certain, no single latent variable can be labeled as *overall,* or *general, performance.* Overall performance is simply a sum score of whatever measures are at hand. If overall performance is generated by a single rating scale labeled *overall performance,* then the

rater must compute the sum score in his or her head, by whatever personal calculus he or she chooses to use, which may or may not be in conscious awareness. What about the general factor that emerges from the covariance matrix of virtually any set of observed performance scores after controlling for method variance (e.g., Viswesvaran, Schmidt, & Ones, 2005)? Such a factor could arise because trait determinants such as general mental ability and conscientiousness contribute to individual differences in virtually all performance measures. Consequently, if one believes the general factor is a latent variable, then it is reasonable to assert that a set of performance measures simply constitutes another measure of general mental ability. General mental ability is the latent variable. No one has yet given a substantive specification of the general factor in performance content terms. It is always specified as a sum score of specific dimensions.

After reviewing all extant research on performance as a construct, Campbell (2012) has argued for an eight-factor structure (discussed earlier in this chapter) that is invariant across work roles, organizational levels, and type of organizations. The status of each of the factors as a latent variable is a mixed bag. Certainly, the technical performance factor does not represent a latent variable. There is always a technical factor, but it must always be specified as a sum score of assessed performance levels on the specific technical responsibilities of the work role, and it might need to be summed over days, weeks, or years. In contrast, a case can be made, with varying degrees of empirical justification, for the latent variable status of the two subfactors of communication, for the Initiative–Effort factor, and for the subfactors of Counterproductive Work Behavior. Campbell (2012) also argued that the subfactors of leadership and management (shown in Exhibits 22.1 and 22.2) have appeared again and again in leadership research using a variety of measures, and it is reasonable to assert that they represent latent variables of performance. A recent integrative review and meta-analysis of research on trait and behavioral leadership models (DeRue, Nahrgang, Wellman, & Humphrey, 2011) is consistent with this view.

High-Stakes Assessment

As is the case for many other subfields, I/O psychology must deal with the assessment complexities of high-stakes testing. Selection for a job, for promotion, and for entry into educational or training programs are indeed high-stakes decisions. They make up a large and critical segment of the research and practice landscape in I/O psychology, and they significantly influence the lives of tens of millions of people. The complexities are intensified enormously by advances in digital technology and by the ethical, legal, and political environments that influence such decision making.

Each of these testing environment complexities (i.e., technological, ethical, legal, and political) has generated its own literature (cf. Farr & Tippins, 2010; Outtz, 2010). The issues include how to deal with unproctored Internet testing; what feedback to provide to test takers; determining the presence or absence of test bias; the currency of federal guidelines; the ethical responsibilities of I/O psychologists; and the efficacy of using changes in standardized test scores to evaluate the value added by teachers, school systems, and universities. Again, these high-stakes issues are simply part of the I/O psychology assessment landscape, and the field must deal with them as thoroughly and as directly as it can.

SOME FINAL (AT LAST) REMARKS

The basic theme of this chapter is the assertion that assessment in I/O psychology is very, very complex. *Complexity* refers to the sheer number of variables across the dependent, independent, and situational variable spectrums; the multidimensional nature of both the latent and the observed structures for each variable; the difficulties involved in developing the substantive specifications for each dimension and their covariance structures; the multiplicity of assessment purposes; the multiplicity of assessment methods; and the intense interaction between science and practice. The scientist–practitioner model still dominates, and that opens the door to the marketplace, high-stakes decision making, the individual versus organizational perspectives, and the attendant value judgments that elicit professional guidelines, governmental rule making, and litigation precedents, all of which have important and complex implications for assessment.

The future will become even more complex. The world of work itself becomes ever more complicated as technology, globalization, population growth, climate science, and competing political ideologies contrive to shape it. These forces will shape psychological assessment as well. For example, Embretson (2004) forecast measurement technologies for the 21st century that could barely be imagined a decade ago, and the assessment methods of neuroscience are now being adapted to focus on the neural antecedents of work performance (Parasuraman, 2011). Will future graduate training in I/O psychology include becoming familiar with neuroimaging methods (e.g., functional magnetic resonance imaging, event-related potential, and magnetoencephalography)? The short answer is yes. Identity crises (Ryan & Ford, 2010) aside, it is an interesting and intense time in I/O psychology. It will become even more so in the future, and I/O psychologists have much to contribute to the future of both science and practice.

To deal with this complexity more effectively, this chapter made the following basic points. First, given a particular assessment domain of interest, it is imperative to specify its constructs as completely and as carefully as possible and model its covariance structure as precisely as possible. If a new construct is proposed, the ways in which it fits into existing structures, or does not fit, should be specified. It is not in the best interests of research and practice to invent new labels for existing variables and imply that something new and different is being assessed or to propose new variables and let them float above the marketplace without specification and an evidence base. This is not an argument for never investigating anything new. It is an argument for careful specification and research-based assessment.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3–27. doi:10.1037/0033-2909.102.1.3

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288–318. doi:10.1037/0096-3445.117.3.288

Ackerman, P. L. (2000). Domain-specific knowledge as the "dark matter" of adult intelligence: Gf/Gc, personality and interest correlates. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 55,* 69–84. doi:10.1093/geronb/55.2.P69

Ackerman, P. L., & Rolfhus, E. L. (1999). The locus of adult intelligence: Knowledge, abilities, and non-ability traits. *Psychology and Aging, 14*, 314–330. doi:10.1037/0882-7974.14.2.314

Albrecht, S. L. (Ed.). (2010). *Handbook of employee engagement: Perspectives, issues, research and practice.* Glos, England: Edward Elgar.

Alderfer, C. P. (1969). An empirical test of a new theory of human needs. *Organizational Behavior and Human Performance, 4,* 142–175. doi:10.1016/0030-5073(69)90004-X

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*, 192–210. doi:10.1037/0033-295X.94.2.192

Arad, S., Hanson, M., & Schneider, R. J. (1999). Organizational context. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jenneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 147–174). Washington, DC: American Psychological Association. doi:10.1037/10313-009

Austin, J. R., & Bartunek, J. M. (2003). Theories and practices of organizational development. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 309–332). Hoboken, NJ: Wiley.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77*, 836–874. doi:10.1037/0021-9010.77.6.836

Bakker, A. B., & Leiter, M. P. (Eds.). (2010). *Work engagement: A handbook of essential theory and research.* New York, NY: Psychology Press.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*, 122–147. doi:10.1037/0003-066X.37.2.122

Banta, T. W. (2008). Editor's notes: Trying to clothe the emperor. *Assessment Update, 20,* 3–4, 15–16.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54*, 462–479. doi:10.1037/0003-066X.54.7.462

Bar-On, R. (1997). *Bar-On Emotional Quotient Inventory: A measure of emotional intelligence.* Toronto, Ontario, Canada: Multi-Health Systems.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., & Mount, M. K. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance, 18,* 359–372. doi:10.1207/s15327043hup1804_3

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90,* 1185–1203. doi:10.1037/0021-9010.90.6.1185

Baum, J. R., Bird, B. J., & Singh, S. (2011). The practical intelligence of entrepreneurs: Antecedents and a link with new venture growth. *Personnel Psychology, 64,* 397–425. doi:10.1111/j.1744-6570.2011.01214.x

Beal, D. J., Weiss, H. M., Barros, E., & MacDermid, S. M. (2005). An episodic process model of affective influences on performance. *Journal of Applied Psychology, 90*, 1054–1068. doi:10.1037/0021-9010.90.6.1054

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*, 349–360. doi:10.1037/0021-9010.85.3.349

Bennett, W., Lance, C. E., & Woehr, D. J. (Eds.). (2006). *Performance measurement: Current perspectives and future challenges*. Mahwah, NJ: Erlbaum.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92*, 410–424. doi:10.1037/0021-9010.92.2.410

Boisot, M., & McKelvey, B. (2010). Integrating modernist and postmodernist perspectives on organizations: A complexity science bridge. *Academy of Management Review, 35*, 415–433. doi:10.5465/AMR.2010.51142028

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. *Organizational Behavior and Human Decision Processes, 40*, 307–322. doi:10.1016/0749-5978(87)90018-5

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*, 1–21. doi:10.1207/s15327043hup0601_1

Borman, W. C., Buck, D. E., Hanson, M. S., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86,* 965–973. doi:10.1037/0021-9010.86.5.965

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco, CA: Jossey-Bass.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory and Psychology, 14*, 105–120. doi:10.1177/0959354304040200

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219. doi:10.1037/0033-295X.110.2.203

Bowers, D. G., & Seashore, W. E. (1966). Predicting organizational effectiveness with a four factor theory of leadership. *Administrative Science Quarterly, 11*, 238–263. doi:10.2307/2391247

Brown, S. P. (1996). A meta-analysis and review of organizational research on job involvement. *Psychological Bulletin, 120*, 235–255. doi:10.1037/0033-2909.120.2.235

Brown, S. P., & Leigh, T. W. (1996). A new look at psychological climate and its relationship to job involvement, effort, and performance. *Journal of Applied Psychology, 81*, 358–368. doi:10.1037/0021-9010.81.4.358

Cameron, K. S., & Quinn, R. E. (1999). *Diagnosing and changing organizational culture: Based on the competing values framework*. Reading, MA: Addison-Wesley.

Campbell, J. P. (1977). On the nature of organizational effectiveness. In P. S. Goodman & J. M. Pennings (Eds.), *New perspectives on organizational effectiveness* (pp. 13–55). San Francisco, CA: Jossey-Bass.

Campbell, J. P. (2007). Profiting from history. In L. L. Koppes, P. W. Thayer, A. J. Vinchur, & E. Salas (Eds.), *Historical perspectives in industrial and organizational psychology* (pp. 441–457). Mahwah, NJ: Erlbaum.

Campbell, J. P. (2012). Behavior, performance, and effectiveness in the 21st century. In S. Kozlowski (Ed.), *Oxford handbook of industrial and organizational psychology* (pp. 159–194). New York, NY: Oxford University Press.

Campbell, J. P., & Knapp, D. (2001). *Exploring the limits of personnel selection and classification*. Hillsdale, NJ: Erlbaum.

Campbell, J. P., & Kuncel, N. R. (2001). Individual and team training. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of work and organizational psychology* (pp. 278–312). London, England: Blackwell.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt &

W. C. Borman (Eds.), *Frontiers in industrial/organizational psychology: Personnel selection and classification* (pp. 35–71). San Francisco, CA: Jossey-Bass.

Campion, M. S., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64,* 225–262. doi:10.1111/j.1744-6570.2010.01207.x

Carlson, K. D. (1997). *Impact of instructional strategy on training effectiveness.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511571312

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about 10 broad factors. In N. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). Amsterdam, the Netherlands: Pergamon Press.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston, MA: Houghton Mifflin.

Chan, D. (2010). Values, styles, and motivational constructs. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 321–337). New York, NY: Routledge.

Christian, M. S., Garza, A. S., & Slaughter, J. E. (2011). Work engagement: A quantitative review and test of its relations with task and contextual performance. *Personnel Psychology, 64,* 89–136. doi:10.1111/j.1744-6570.2010.01203.x

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Erlbaum.

Cleveland, J. N., & Colella, A. (2010). Employee work-related health, stress, and safety. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 531–550). New York, NY: Routledge.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology, 86*, 386–400. doi:10.1037/0021-9010.86.3.386

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology, 86*, 425–445. doi:10.1037/0021-9010.86.3.425

Conway, J. M. (1998). Understanding method variance in multitrait–multirater performance appraisal matrices: Examples using general impressions and interpersonal effect as measured method factors. *Human Performance, 11*, 29–55. doi:10.1207/s15327043hup1101_2

Conway, J. M., & Huffcut, A. L. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331–360. doi:10.1207/s15327043hup1004_2

Cooke, R. A., & Rousseau, D. M. (1988). Behavioral norms and expectations: A quantitative approach to the assessment of organizational culture. *Group and Organization Management, 13*, 245–273. doi:10.1177/105960118801300302

Cooper, C. L., Dewe, P., & O'Driscoll, M. P. (2001). *Organizational stress: A review and critique of theory, research, and applications.* Thousand Oaks, CA: Sage.

Costa, P. T., Jr., & McRae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI–R) and NEO Five-Factor Inventory (NEO FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Crede, M. (2006). *Job attitude and job evaluation: Examining construct-measurement discrepancies.* Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*, 1241–1255.

Davenport, R. (2006). Eliminate the skills gap. *Training and Development, 60*, 27–32.

Dawis, R. V., Dohm, T. E., Lofquist, L. H., Chartrand, J. M., & Due, A. M. (1987). *Minnesota Occupational Classification System III: A psychological taxonomy of work.* Minneapolis: University of Minnesota, Department of Psychology, Vocational Psychology Research.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment.* Minneapolis: University of Minnesota Press.

Deadrick, D. L., Bennett, N., & Russell, C. J. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management, 23*, 745–757. doi:10.1177/014920639702300603

Denison, D. R. (1996). What is the difference between organizational culture and organizational climate? A native's point of view on a decade of paradigm wars. *Academy of Management Review, 21*, 619–654.

DeRue, D. S., Nahrgang, J. D., Wellman, N., & Humphrey, S. E. (2011). Trait and behavioral theories of leadership: An integration and meta-analytic test of their relative validity. *Personnel Psychology, 64*, 7–52. doi:10.1111/j.1744-6570.2010.01201.x

DeShon, R. P., & Gillespie, J. Z. (2005). A motivated action theory account of goal orientation. *Journal of Applied Psychology, 90,* 1096–1127. doi:10.1037/0021-9010.90.6.1096

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology, 91,* 1138–1151. doi:10.1037/0022-3514.91.6.1138

DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the Big Five. *Psychological Science, 21*, 820–828. doi:10.1177/0956797610370159

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93*, 880–896. doi:10.1037/0022-3514.93.5.880

Dilchert, S. (2008). *Measurement and prediction of creativity at work*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 465–476. doi:10.1111/j.1754-9434.2010.01273.x

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–1048. doi:10.1037/0003-066X.41.10.1040

Eagley, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. New York, NY: Wadsworth.

Edwards, J. E., & Rothbard, N. P. (2000). Mechanisms linking work and family: Clarifying the relationship between work and family constructs. *Academy of Management Review, 25,* 178–199.

Edwards, J. R., Scully, J. S., & Bartek, M. D. (1999). The measurement of work: Hierarchical representation of the multimethod job design questionnaire. *Personnel Psychology, 52*, 305–334. doi:10.1111/j.1744-6570.1999.tb00163.x

Elliot, A. J., & Thrash, T. M. (2002). Approach–avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology, 82*, 804–818. doi:10.1037/0022-3514.82.5.804

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*, 5–12. doi:10.1037/0022-3514.54.1.5

Embretson, S. E. (2004). The second century of ability testing: Some new predictions and speculations. *Measurement, 2*, 1–32.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61*, 50–55. doi:10.1037/0003-066X.61.1.50

Embretson, S. E., & Reise, S. P. (Eds.). (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*, 44–48.

Ericsson, K. S., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. New York, NY: Cambridge University Press.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. doi:10.1146/annurev.psych.59.103006.093629

Ewell, P. T. (1991). To capture the ineffable: New forms of assessment in higher education. *Review of Research in Education, 17*, 75–125.

Eysenck, H. J. (1967). *The biological basis of personality*. Springfield, IL: Charles C Thomas.

Farr, J. L., & Tippins, N. T. (2010). *Handbook of employee selection*. New York, NY: Routledge.

Ferris, L. D., Rosen, C. R., Johnson, R. E., Brown, D. J., Risavy, S. D., & Heller, D. (2011). Approach or avoidance (or both)? Integrating core self-evaluations within an approach/avoidance framework. *Personnel Psychology, 64,* 137–161. doi:10.1111/j.1744-6570.2010.01204.x

Fleishman, E. A. (1964). *The structure and measurement of physical fitness*. Englewood Cliffs, NJ: Prentice Hall.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks*. New York, NY: Academic Press.

Fleishman, E. A., & Reilly, M. E. (1992). *Handbook of human abilities: Definitions, measurements, and job task requirements*. Bethesda, MD: Management Research Institute.

Gable, S. L., Reis, H. T., & Elliot, A. J. (2003). Evidence for bivariate systems: An empirical test of appetition and aversion across domains. *Journal of Research in Personality, 37*, 349–372. doi:10.1016/S0092-6566(02)00580-9

Galagan, P. (2010). *Bridging the skills gap: New factors compound the growing skills shortage*. Alexandria, VA: American Society for Training and Development.

Gebhardt, D. L., & Baker, T. A. (2010). Physical performance tests. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 277–298). New York, NY: Routledge.

George, J. M. (2007). Creativity in organizations. *Academy of Management Annals, 1*, 439–477. doi:10.1080/078559814

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York, NY: Wiley.

Gollwitzer, P. M., Sheeran, P., Trotschel, R., & Webb, T. L. (2011). Self-regulation of priming effects on behavior. *Psychological Science, 22*, 901–907. doi:10.1177/0956797611411586

Goodman, P. S., Devadas, R., & Griffith-Hughson, T. L. (1988). Groups and productivity: Analyzing the effectiveness of self-management teams. In J. P. Campbell, R. J. Campbell, & Associates (Eds.), *Productivity in organizations: New perspectives from industrial and organizational psychology* (pp. 295–327). San Francisco: Jossey-Bass.

Gottfredson, L. S. (2003). Dissenting practical intelligence theory: Its claims and evidence. *Intelligence, 31*, 343–397. doi:10.1016/S0160-2896(02)00085-5

Greenhaus, J. H., & Powell, G. N. (2006). When work and family are allies: A theory of work-family enrichment. *Academy of Management Review, 31,* 72–92. doi:10.5465/AMR.1985.4277352

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Updated moderator tests, and research implications for the next millennium. *Journal of Management, 26*, 463–488. doi:10.1177/014920630002600305

Griffin, M. S., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50*, 327–347. doi:10.5465/AMJ.2007.24634438

Gruys, M. L., & Sackett, P. R. (2003). Investigating the dimensionality of counterproductive work behavior. *International Journal of Selection and Assessment, 11*, 30–42. doi:10.1111/1468-2389.00224

Grzywacz, J. G., & Carlson, D. S. (2007). Conceptualizing work-family balance: Implications for practice and research. *Advances in Developing Human Resources, 9,* 455–471. doi:10.1177/1523422307305487

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135–164. doi:10.1111/j.1744-6570.1965.tb00273.x

Hackman, J. R. (1992). Group influences on individuals in organizations. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 199–267). Palo Alto, CA: Consulting Psychologists Press.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*, 250–279. doi:10.1016/0030-5073(76)90016-7

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory: Applications and technical guide.* Stanford, CA: Stanford University Press.

Harnois, G., & Gabriel, P. (2000). *Mental health and work: Impact, issues and good practices.* Geneva, Switzerland: International Labour Organisation.

Herzberg, F. (1959). *The motivation to work.* New York, NY: Wiley.

Hesketh, B., & Neal, A. (1999). Technology and performance. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 21–55). San Francisco, CA: Jossey-Bass.

Hobfoll, S. E. (1998). *Stress, culture, and community: The psychology and physiology of stress.* New York, NY: Plenum Press.

Hofmann, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77*, 185–195. doi:10.1037/0021-9010.77.2.185

Hogan, J. (1991). Structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76*, 495–507. doi:10.1037/0021-9010.76.4.495

Hogan, R., & Kaiser, R. B. (2010). Personality. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 81–108). San Francisco, CA: Jossey-Bass.

Holland, J. L. (1994). *The Self-Directed Search: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Hoppock, R. (1935). *Job satisfaction.* New York, NY: Harper.

Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 61–116). New York, NY: Freeman.

Hough, L., & Dilchert, S. (2010). Personality: Its measurement and validity for employee selection. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 299–319). New York, NY: Routledge.

Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology* (pp. 233–277). Thousand Oaks, CA: Sage.

Hulin, C. L., & Judge, T. A. (2003). Job attitudes. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 255–276). Hoboken, NJ: Wiley.

Hunt, J. G. (1999). Transformational/charismatic leadership's transformation of the field: An historical essay. *Leadership Quarterly, 10*, 129–144. doi:10.1016/S1048-9843(99)00015-6

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology, 56*, 517–543. doi:10.1146/annurev.psych.56.091103.070250

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin, 81*, 1096–1112. doi:10.1037/h0037511

Johnson, P., & Cassell, C. (2001). Epistemology and work psychology: New agendas. *Journal of Occupational and Organizational Psychology, 74*, 125–143. doi:10.1348/096317901167280

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 393–416. doi:10.1016/j.intell.2004.12.002

Johnson, W., Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 *g*: Consistent result from five test batteries. *Intelligence, 36*, 81–95. doi:10.1016/j.intell.2007.06.001

Judge, T. A., & Bono, J. E. (2001). A rose by any other name: Are self-esteem, generalized self-efficacy, neuroticism, and locus of control indicators of a common construct? In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace* (pp. 93–118). Washington, DC: American Psychological Association. doi:10.1037/10434-004

Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job life satisfaction: The role of core evaluation. *Journal of Applied Psychology, 83*, 17–34. doi:10.1037/0021-9010.83.1.17

Judge, T. A., Van Vianen, A. E. M., & DePater, I. E. (2004). Emotional stability, core self-evaluations, and job outcomes: A review of the evidence and an agenda for future research. *Human Performance, 17*, 325–346. doi:10.1207/s15327043hup1703_4

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515–526. doi:10.1037/a0016755

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291. doi:10.2307/1914185

Kanfer, R., Chen, G., & Pritchard, R. (Eds.). (2008). *Work motivation: Past, present, and future*. New York, NY: Taylor & Francis.

Kanfer, R., & Heggestad, E. D. (1997). Motivational traits and skills: A person-centered approach to work motivation. *Research in Organizational Behavior, 19*, 1–56.

Kanungo, R. N. (1982). Measurement of job and work involvement. *Journal of Applied Psychology, 67*, 341–349. doi:10.1037/0021-9010.67.3.341

Katzell, R. A., & Guzzo, R. A. (1983). Psychological approaches to productivity improvement. *American Psychologist, 38*, 468–472. doi:10.1037/0003-066X.38.4.468

Kelloway, E. K., Loughlin, C., Barling, J., & Nault, A. (2002). Self-reported counterproductive behaviors and organizational citizenship behaviors: Separate but related constructs. *International Journal of Selection and Assessment, 10*, 143–151. doi:10.1111/1468-2389.00201

Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review, 32*, 511–525. doi:10.1177/0193841X08325948

Kline, P. (1997). Commentary on Michell, quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 358–387. doi:10.1111/j.2044-8295.1997.tb02642.x

Koppes, L. L., Thayer, P. W., Vinchur, A. J., & Salas, E. (Eds.). (2007). *Historical perspectives in industrial and organizational psychology*. Mahwah, NJ: Erlbaum.

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7*, 77–124.

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology, 8*, 65–77. doi:10.1111/j.1744-6570.1955.tb01189.x

Landy, F. J. (2005). Some historical and scientific issues related to research on emotional intelligence. *Journal of Organizational Behavior, 26*, 411–424. doi:10.1002/job.317

Levi, L., & Lunde-Jensen, P. (1996). *A model for assessing the costs of stressors at national level: Socio-economic costs of work stress in two EU member states*. Dublin, Ireland: European Foundation for the Improvement of Living and Working Conditions.

Lewin, K. (1951). *Field theory in social science*. New York, NY: Harper & Row.

Liberman, V. (2011). Why your people can't do what you need them to do. *Conference Board Review, Winter,* 1–8.

Lievens, F., & Chan, D. (2010). Practical intelligence, emotional intelligence, and social intelligence. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 339–359). New York, NY: Routledge.

Likert, R. (1932). The method of constructing an attitude scale. *Archives de Psychologie, 140*, 44–53.

Litwin, G. H., & Stringer, R. A. (1968). *Motivation and organizational climate*. Boston, MA: Harvard University Press.

Liu, O. L. (2011). Value-added assessment in higher education: A comparison of two methods. *Higher Education, 61*, 445–461. doi:10.1007/s10734-010-9340-8

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*, 705–717. doi:10.1037/0003-066X.57.9.705

Lodahl, T. M., & Kejner, M. (1965). The definition and measurement of job involvement. *Journal of Applied Psychology, 49,* 24–33. doi:10.1037/h0021692

Lord, R. G., Diefendorff, J. M., Schmidt, A. M., & Hall, R. J. (2010). Self-regulation at work. *Annual Review of Psychology, 61*, 543–568. doi:10.1146/annurev.psych.093008.100314

Lubinski, D. (2010). Neglected aspects and truncated appraisals in vocational counseling: Interpreting the interest-efficacy association from a broader perspective: Comment on Armstrong and Vogel (2009). *Journal of Counseling Psychology, 57,* 226–238. doi:10.1037/a0019163

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new scale type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27. doi:10.1016/0022-2496(64)90015-X

Lykken, D. T. (1999). *Happiness: What studies on twins show us about nature, nurture, and the happiness set point.* New York, NY: Golden Books.

Macey, W. H., & Schneider, B. (2008). The meaning of employee engagement. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 3–30. doi:10.1111/j.1754-9434.2007.0002.x

Macey, W. H., Schneider, B., Barbera, K., & Young, S. A. (2009). *Employee engagement: Tools for analysis, practice, and competitive advantage.* London, England: Blackwell.

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology, 90,* 710–730. doi:10.1037/0021-9010.90.4.710

Maertz, C. P., & Campion, M. A. (2004). Profiles in quitting: Integrating process and content turnover theory. *Academy of Management Journal, 47,* 566–582. doi:10.2307/20159602

Marcus, B., Schuler, H., Quell, P., & Humpfner, G. (2002). Measuring counterproductivity: Development and initial validation of a German self-report questionnaire. *International Journal of Selection and Assessment, 10*, 18–35. doi:10.1111/1468-2389.00191

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. *Journal of Personality and Social Psychology, 88,* 139–157. doi:10.1037/0022-3514.88.1.139

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*, 370–396. doi:10.1037/h0054346

Masson, R. C., Royal, M. A., Agnew, T. G., & Fine, S. (2008). Leveraging employee engagement: The practical implications. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 56–59. doi:10.1111/j.1754-9434.2007.00009.x

Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997–2007: A review of recent advancements and a glimpse into the future. *Journal of Management, 34*, 410–476. doi:10.1177/0149206308316061

McClelland, D. C. (1985). How motives, skills, and values determine what people do. *American Psychologist, 40,* 812–825. doi:10.1037/0003-066X.40.7.812

McPhail, S. M. (Ed.). (2007). *Alternative validation strategies: Developing new and leveraging existing validity evidence.* San Francisco, CA: Jossey-Bass.

Meyer, H. H. (2007). Influence of formal and informal organizations on the development of I-O psychology. In L. L. Koppes, P. W. Thayer, A. J. Vinchur, & E. Salas (Eds.), *Historical perspectives in industrial and organizational psychology* (pp. 139–168). Mahwah, NJ: Erlbaum.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511490040

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology, 10*, 639–667. doi:10.1177/0959354300105004

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6*, 7–24.

Miles, D. E., Borman, W. C., Spector, P. E., & Fox, S. (2002). Building an integrative model of extra role work behaviors: A comparison of counterproductive work behavior with organizational citizenship behavior. *International Journal of Selection and Assessment, 10*, 51–57. doi:10.1111/1468-2389.00193

Miner, J. B. (1977). *Motivation to manage: A ten-year update on the "studies in management education" research.* Atlanta, GA: Organizational Measurement Systems Press.

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement, 6*, 124.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215. doi:10.1007/BF02295283

Mitchell, T. R., & Daniels, D. (2003). Motivation. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 225–254). Hoboken, NJ: Wiley.

Mitchell, T. R., & Lee, T. W. (2001). The unfolding model of voluntary turnover and job embeddedness: Foundations for a comprehensive theory of attachment. In B. Staw & R. Sutton (Eds.), *Research in organizational behavior* (Vol. 23, pp. 189–246). Stamford, CT: JAI Press.

Morgeson, F. P., & Campion, M. E. (2003). Work design. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 423–452). Hoboken, NJ: Wiley.

Morgeson, F. P., Campion, M. S., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683–729. doi:10.1111/j.1744-6570.2007.00089.x

Mumford, M. D., Baughman, W. S., Supinski, E. P., Costanza, D. P., & Threlfall, K. V. (1996). Process-based measures of creative problem solving skills: Overall prediction. *Creativity Research Journal, 9*, 63–76. doi:10.1207/s15326934crj0901_6

Murphy, K. R. (1989a). Dimensions of job performance. In R. Dillon & J. Pelligrino (Eds.), *Testing: Applied and theoretical perspectives* (pp. 218–247). New York, NY: Praeger.

Murphy, K. R. (1989b). Is the relationship between cognitive ability and job performance stable over time? *Human Performance, 2*, 183–200. doi:10.1207/s15327043hup0203_3

Murphy, K. R. (Ed.). (2006). *A critique of emotional intelligence: What are the problems and how can they be fixed?* Mahwah, NJ: Erlbaum.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

Murray, H. A. (1938). *Explorations in personality*. New York, NY: Oxford University Press.

Myers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1993). The dimensions of human physical performance: Factor analysis of strength, stamina, flexibility, and body composition measures. *Human Performance, 6*, 309–344. doi:10.1207/s15327043hup0604_2

National Institute of Occupational Safety and Health. (1999). *Stress … at work* (DHHS Publication No. 99–101). Cincinnati, OH: Author.

Neuman, J. H. (2004). Injustice, stress, and aggression in organizations. In R. W. Griffin & A. M. O'Leary-Kelly (Eds.), *The dark side of organizational behavior* (pp. 62–102). San Francisco, CA: Jossey-Bass.

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96,* 762–773. doi:10.1037/a0021832

Olson, A. M. (2000). *A theory and taxonomy of individual team member performance*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60,* 995–1027. doi:10.1111/j.1744-6570.2007.00099.x

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2010). Cognitive abilities. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 255–275). New York, NY: Routledge.

Ones, D. S., & Viswesvaran, C. (2003). Personality and counterproductive work behaviors. In M. Koslowsky, S. Stashevsky, & A. Sagie (Eds.), *Misbehavior and dysfunctional attitudes in organizations* (pp. 211–249). Hampshire, England: Palgrave Macmillan.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington Books.

Ostroff, C. (1993). The effects of climate and personal influences on individual behavior and attitudes in organizations. *Organizational Behavior and Human Decision Processes, 56*, 56–90. doi:10.1006/obhd.1993.1045

Ostroff, C., Kinicki, A. J., & Tamkins, M. (2003). Organizational culture and climate. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 565–594). Hoboken, NJ: Wiley.

Outtz, J. L. (2010). Addressing the flaws in our assessment decisions. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 711–727). San Francisco, CA: Jossey-Bass.

Parasuraman, R. (2011). Neuroergonomics: Brain, cognition, and performance at work. *Current Directions in Psychological Science, 20,* 181–186. doi:10.1177/0963721411409176

Parry, S. B. (1996). The quest for competencies. *Training, 33*, 48–54.

Paul, R., & Elder, L. (2006). *Critical thinking tools for taking charge of your learning and your life.* Upper Saddle River, NJ: Prentice Hall.

Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology, 92,* 128–150. doi:10.1037/0021-9010.92.1.128

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association. doi:10.1037/10313-000

Ployhart, R. E., & Bliese, P. D. (2006). Individual adaptability (I-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. In E. Salas (Ed.), *Advances in human performance and cognitive engineering research* (Vol. 6, pp. 3–39). Oxford, England: Emerald Group.

Ployhart, R. E., & Hakel, M. D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51*, 859–901. doi:10.1111/j.1744-6570.1998.tb00744.x

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., & Lee, J. Y. (2003). The mismeasure of man(agement) and its implications for leadership research. *Leadership Quarterly, 14*, 615–656. doi:10.1016/j.leaqua.2003.08.002

Pritchard, R. D., Holling, H., Lammers, F., & Clark, B. D. (Eds.). (2002). *Improving organizational performance with the productivity measurement and enhancement system: An international collaboration.* Huntington, NY: Nova Science.

Pulakos, E. D., Arad, S., Donovan, M. S., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*, 612–624. doi:10.1037/0021-9010.85.4.612

Pulakos, E. D., & O'Leary, R. S. (2010). Defining and measuring results of workplace behavior. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 513–529). New York, NY: Routledge.

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9–49). New York, NY: Routledge.

Quinn, R. W., & Rohrbaugh, J. (1983). A spatial model of effectiveness criteria: Towards a competing values approach to organizational analysis. *Management Science, 29*, 363–377. doi:10.1287/mnsc.29.3.363

Reb, J., & Cropanzano, R. (2007). Evaluating dynamic performance: The influence of salient gestalt characteristics on performance ratings. *Journal of Applied Psychology, 92*, 490–499. doi:10.1037/0021-9010.92.2.490

Robert, G., & Hockey, J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology, 45*, 73–93. doi:10.1016/S0301-0511(96)05223-4

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38*, 555–572. doi:10.2307/256693

Rosse, R. L., Campbell, J. P., & Peterson, N. G. (2001). Personnel classification and differential job assignments: Estimating classification gains. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits of personnel selection and classification* (pp. 453–506). Hillsdale, NJ: Erlbaum.

Runco, M. A. (2004). Creativity. *Annual Review of Psychology, 55*, 657–687. doi:10.1146/annurev.psych.55.090902.141502

Ryan, A. M., & Ford, K. J. (2010). Organizational psychology and the tipping point of professional identity. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 241–258. doi:10.1111/j.1754-9434.2010.01233.x

Sackett, P. R., & Laczo, R. M. (2003). Job and work analysis. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 21–37). Hoboken, NJ: Wiley.

Salas, E., Rosen, M. S., & DiazGranados, D. (2010). Expertise-based intuition and decision making in organizations. *Journal of Management, 36,* 941–973. doi:10.1177/0149206309350084

Salovey, P., & Mayer, J. D. (1989–1990). Emotional intelligence. *Imagination, Cognition and Personality, 9*, 185–211.

Secretary's Commission on Achieving Necessary Skills. (1999). *Skills and tasks for jobs.* Washington, DC: U.S. Department of Labor.

Schippman, J. S. (2010). Competencies, job analysis, and the next generation of modeling. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 197–231). San Francisco, CA: Jossey-Bass.

Schippman, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., … Sanchez, J. I. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703–740. doi:10.1111/j.1744-6570.2000.tb00220.x

Schmidt, A. M., Dolis, C. M., & Tolli, A. P. (2009). A matter of time: Individual differences, contextual dynamics, and goal progress effects on multiple-goal self-regulation. *Journal of Applied Psychology, 94*, 692–709. doi:10.1037/a0015012

Schmidt, F., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274. doi:10.1037/0033-2909.124.2.262

Schneider, B. (1990). The climate for service: An application of the climate construct. In B. Schneider (Ed.),

Organizational climate and culture (pp. 383–412). San Francisco, CA: Jossey-Bass.

Schraagen, J. M., Chipman, S. F., & Shalin, V. (Eds.). (2000). *Cognitive task analysis.* Mahwah, NJ: Erlbaum.

Schwartz, S. H., & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross cultural replications. *Journal of Personality and Social Psychology, 58*, 878–891. doi:10.1037/0022-3514.58.5.878

Scott, J. C., & Pearlman, K. (2010). Assessment for organizational change: Mergers, restructuring, and downsizing. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 533–575). San Francisco, CA: Jossey-Bass.

Scott, J. C., & Reynolds, D. H. (Eds.). (2010). *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent.* San Francisco, CA: Jossey-Bass.

Selye, H. (1975). Confusion and controversy in the stress field. *Journal of Human Stress, 1*, 37–44. doi:10.1080/0097840X.1975.9940406

Simon, H. A. (1992). What is an explanation of behavior? *Psychological Science, 3*, 150–161. doi:10.1111/j.1467-9280.1992.tb00017.x

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago, IL: Rand McNally.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures.* Bowling Green, OH: Author.

Sonnentag, S., & Frese, M. (2003). Stress in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 453–492). Hoboken, NJ: Wiley.

Sonnentag, S., & Frese, M. (2012). Performance dynamics. In S. Kozlowski (Ed.), *Oxford handbook of industrial and organizational psychology* (pp. 548–575). New York, NY: Oxford University Press.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201–293. doi:10.2307/1412107

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology, 95*, 781–790. doi:10.1037/a0019477

Steedle, J., Kugelmass, H., & Nemeth, A. (2010). What do they measure? Comparing three learning outcomes assessment. *Change: The Magazine of Higher Learning, 42*, 33–37. doi:10.1080/00091383.2010.490491

Steel, P., & Konig, C. (2006). Integrating theories of motivation. *Academy of Management Review, 31,* 889–913. doi:10.5465/AMR.2006.22527462

Sternberg, R. J. (2003). A broad view of intelligence: The theory of successful intelligence. *Consulting Psychology Journal: Practice and Research, 55*, 139–154. doi:10.1037/1061-4087.55.3.139

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50*, 912–927. doi:10.1037/0003-066X.50.11.912

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680. doi:10.1126/science.103.2684.677

Stewart, G. L., & Nandkeolyar, A. K. (2006). Adaptation and intraindividual variation in sales outcomes: Exploring the interactive effects of personality and environmental opportunity. *Personnel Psychology, 59*, 307–332.

Stice, J. E. (Ed.). (1987). *Teaching critical thinking and problem solving abilities.* San Francisco, CA: Jossey-Bass.

Strong, M. H., Jeanneret, P. R., McPhail, S. M., Blakley, B. R., & D'Egidio, E. L. (1999). Work context: Taxonomy and measurement of the work environment. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jenneret, & E. A., Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 127–145). Washington, DC: American Psychological Association. doi:10.1037/10313-008

Sturman, M. C. (2003). Searching for the inverted u-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management, 29*, 609–640.

Sullivan, B. A., & Hansen, J. C. (2004). Mapping associations between interests and personality: Toward a conceptual understanding of individual differences in vocational behavior. *Journal of Counseling Psychology, 51,* 287–298. doi:10.1037/0022-0167.51.3.287

Taras, V., Kirkman, B. L., & Steel, P. (2010). Examining the impact of culture's consequences: A three-decade, multilevel, meta-analytic review of Hofstede's cultural value dimensions. *Journal of Applied Psychology, 95,* 405–439. doi:10.1037/a0018938

Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota, Minneapolis.

Tellegen, A., & Waller, N. (2000). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1, pp. 133–161). Greenwich, CT: JAI Press.

Tetrick, L., Perrewé, P. L., & Griffin, M. (2010). Employee work-related health, stress, and safety. In J. L. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 531–549). New York, NY: Routledge.

Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. A. (2000). Development and content validation of a "hyperdimensional" taxonomy of managerial competence. *Human Performance, 13,* 205–251. doi:10.1207/S15327043HUP1303_1

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554. doi:10.1086/214483

Tippins, N. T., & Hilton, M. L. (Eds.).; Panel to Review the Occupational Informational Network (O*NET), National Research Council. (2010). *A database for a changing economy: Review of the Occupational Information Network (O*NET)*. Washington, DC: National Academies Press.

Trice, H. M., & Beyer, J. M. (1993). *The cultures of work organizations*. Englewood Cliffs, NJ: Prentice Hall.

Unsworth, K. (2001). Unpacking creativity. *Academy of Management Review, 2*, 289–297.

U.S. Office of Personnel Management. (2007). *Delegated examining operations handbook: A guide for federal agency examining offices*. Washington, DC: U.S. Office of Personnel Management.

Van Iddekinge, C. H., Putka, D. J., & Campbell, J. P. (2011). Reconsidering vocational interests for personnel selection: The validity of an interest-based selection test in relation to job knowledge, job performance and continuance intentions. *Journal of Applied Psychology*, *96,* 13–33. doi:10.1037/a0021193

Verbeke, W., Volgering, M., & Hessels, M. (1998). Exploring the conceptual expansion within the field of organizational behavior: Organizational climate and organizational culture. *Journal of Management Studies, 35*, 303–329. doi:10.1111/1467-6486.00095

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131. doi:10.1037/0021-9010.90.1.108

Vosburgh, R. M. (2008). State-trait returns! And one practitioner's request. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 72–73. doi:10.1111/j.1754-9434.2007.00014.x

Vroom, V. (1964). *Work and motivation.* Chichester, England: Wiley.

Warr, P. B. (1987). *Work, unemployment, and mental health.* Oxford, England: Oxford University Press.

Warr, P. B. (1994). A conceptual framework for the study of work and mental health. *Work and Stress, 8*, 84–97. doi:10.1080/02678379408259982

Watson, D., & Clark, L. A. (1993). Behavioral disinhibition versus constraint: A dispositional perspective. In D. M. Wegner & J. W. Pennebaker (Eds.), *Handbook of mental control* (pp. 506–527). New York, NY: Prentice Hall.

Weiss, H. M. (2002). Deconstructing job satisfaction: Separating evaluations, beliefs, and affective experiences. *Human Resource Management Review, 12*, 173–194. doi:10.1016/S1053-4822(02)00045-1

Weiss, H. M., & Rupp, D. E. (2011). Experiencing work: An essay on a person-centric work psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 83–97. doi:10.1111/j.1754-9434.2010.01302.x

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*, 297–333. doi:10.1037/h0040934

Yukl, G. A., Gordon, A., & Taber, T. (2002). A hierarchical taxonomy of leadership behavior: Integrating a half century of behavior research. *Journal of Leadership and Organizational Studies, 9*, 15–32. doi:10.1177/107179190200900102

Zedeck, S. (Ed.). (2010). *APA handbook of industrial and organizational psychology*. Washington, DC: American Psychological Association.

Zeidner, J., Johnson, C. D., & Scholarios, D. (1997). Evaluating military selection and classification systems in the multiple job context. *Military Psychology, 9*, 169–186. doi:10.1207/s15327876mp0902_4

Zhou, J., & Shalley, C. E. (2003). Research on employee creativity: A critical review and directions for future research. In J. J. Martocchio & G. R. Ferris (Eds.), *Research in personnel and human resource management* (Vol. 22, pp. 165–217). Oxford, England: Elsevier Science.

Zyphur, M. J., Chaturvedi, S., & Arvey, R. (2008). Job performance over time is a function of latent trajectories and previous performance. *Journal of Applied Psychology, 93*, 217–224. doi:10.1037/0021-9010.93.1.217

CHAPTER 23

WORK ANALYSIS FOR ASSESSMENT

Juan I. Sanchez and Edward L. Levine

The purpose of this chapter is to review extant research and practices concerning the role that job analysis plays in the assessment process. *Job analysis* is defined via a combination of two definitions adapted from Brannick, Levine, and Morgeson (2007) and Sanchez and Levine (2012). Job analysis is made up of a set of systematic methods aimed at explaining what people do at work and the context in which they do it, understanding the essential nature and meaning of their role in an organization, and elucidating the human attributes needed to carry out their role. Although the target of the analysis is often a set of positions that together are labeled a *job,* job analysis need not be confined by job boundaries but may instead focus on segments of the job, teams, and the broader role enacted by people in organizations. To signify this broader focus, the more encompassing term *work analysis* has been proposed in lieu of *job analysis* (Sanchez, 1994; Sanchez & Levine, 1999, 2001). *Work analysis* has been the term of choice in recent reviews of the literature (Morgeson & Dierdorff, 2011; Sanchez & Levine, 2012). Both terms—*job analysis* and *work analysis*—are used interchangeably in this chapter.

Essentially, job analysis has been used since its inception to ensure that individual assessments target those behaviors and attributes required for performance of a job or group of jobs, as opposed to arbitrary or irrelevant behaviors and attributes (Münsterberg, 1913; Stern, 1911). For instance, the preferred method for ensuring the job relatedness of licensure and credentialing assessments for a given occupation is to include a job analysis as part of the assessment development (Raymond, 2001; Smith & Hambleton, 1990). The effectiveness of virtually all human resource management practices, including selection, training performance management, career planning, team performance enhancement, worker mobility, and deployment of staff, depends on valid assessments (Brannick et al., 2007). It has also been a foundational assumption throughout the history of the field of industrial and organizational psychology that job analysis serves an irrefutable role in ensuring the development of valid assessments. This review of practices and research highlights how job analysis fulfills this role.

As such, the various decisions or inferences that are supported by job analysis—from the determination of important job behaviors and associated personal attributes to the formulation of an assessment plan—are reviewed. Highlighting the purposeful role of job analysis helps overcome the conceptualization of job analysis as merely a methodology, which has emphasized procedural choices such as the choice of sources and methods through which job information should be gathered (Pearlman & Sanchez, 2010; Sackett & Laczo, 2003; Sanchez & Levine, 1999, 2001). This notion detracts from attention more appropriately directed toward the rules through which job-analytic information is used to draw assessment-related inferences. This focus aligns job analysis with the dominant conceptualization of construct validity as being concerned with inferences and their consequences. In contrast to the notion of job analysis as primarily a series of methodological choices, it is proposed that job

DOI: 10.1037/14047-023
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

analysis should play an integral part in the substantive decisions that inform the assessment process, such as the derivation of human attributes to be targeted in the assessment and the establishment of an assessment plan that reflects the relative importance of the various aspects of the job. This emphasis on how job-analytic data should be used to support valid decision making is predicated on the assumption that the frequently observed emphasis on the quality of job-analytic data is misplaced, because excellent data may lead to faulty decisions when data are accompanied by vague or nonexistent decision-making rules. In short, the rules governing the making of inferences derived from job-analytic information are the focus of this chapter, because such inferences constitute the most critical contribution of job analysis to the assessment process.

Conducting job-related assessments makes business sense because it ensures that employees possess the requirements to perform their job successfully, but there is also a legal mandate to stick to job-related information in workplace assessments in the United States. Indeed, the Civil Rights Act of 1964 and subsequent court rulings dealing with equal employment opportunity as defined in that law emphasized the importance of demonstrating the "job-relatedness" and "business necessity" of assessment conducted for employment purposes. According to the Equal Employment Opportunity Commission's *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice, 1978), content-valid work assessments conducted for employment purposes need to be linked to important and critical job behaviors, which are identified through job analysis. When *prima facie* evidence is adduced of adverse impact on members of a social group whose employment rights are protected by law, the burden of proof rests with the employer to show that the requirements evaluated in the assessment are indeed job related or respond to a business necessity. Failure to provide adequate proof indicates that the assessments in question are unlawfully discriminatory. Another statutory requirement for job analysis is present in the Americans with Disabilities Act of 1990. This law requires employers to make reasonable accommodations that would allow qualified disabled workers to perform the essential functions of the job, thereby creating the need to determine essential job behaviors through job analysis (Brannick, Brannick, & Levine, 1992; Mitchell, Alliger, & Morfopoulos, 1997).

Although employers may fulfill these job-relatedness provisions by documenting the empirical association between an assessment and a criterion such as job performance, such criterion-related validation studies are technically unfeasible when sample sizes are small and statistical power is insufficient, which make job analysis–based determinations of content validity an important alternative and one of the only viable ones. For instance, the assessment process to become a firefighter in the city of Dallas, Texas, required climbing a fence 6 feet in height in a prescribed amount of time. Because a higher percentage of women were unable to scale the fence than men within the prescribed time limit, suggesting adverse impact of the assessment on women, the city was asked to demonstrate that climbing this type of fence was job-related. The city presented job analysis evidence indicating that the average fence in the jurisdiction was 6 feet in height and that climbing fences was frequently required of firefighters.

Unfortunately, conducting a job analysis is sometimes seen in the United States as an activity whose sole purpose is to manage the risk of a potential legal challenge, particularly in these litigious times. Ensuring the job relatedness of assessments by linking them to job-analytic information, however, is not, and originally was not, intended to be primarily a litigation tool but rather a way to staff organizations effectively in a manner that reflects the actual requirements of jobs.

The role of job analysis in the assessment process can be viewed as a series of inferences or decisions that link various types of job-analytic information. Typically, the first step involves gathering information on job responsibilities or job tasks. Then, data concerning the context, including the sociophysical environment in which work is carried out, need to be compiled to gain a better understanding of the working conditions under which job responsibilities are discharged. Finally, the attributes or characteristics of the people doing the work are derived from a

combined analysis of the first two domains of job-analytic information, specifically job responsibilities and work context.

The following pages of this chapter are divided in two major sections. The first major section is devoted to a review of how job analysis informs the assessment process, which is embedded within the inferences or "inferential leaps" facilitated by job-analytic information (Gatewood, Feild, & Barrick, 2008): identifying job responsibilities, measuring the work context in which responsibilities are carried out, and deriving the human attributes needed for successful job performance by linking them to previously gathered information on job responsibilities and work context. These linkages are critical for assessment purposes, because they provide the underlying rationale that justifies why certain attributes are needed to perform certain activities under certain conditions. The role that job analysis plays in determining assessment specifications and plans is reviewed; this is a job-analytic application that has largely been neglected in prior reviews of the job analysis literature.

Although this chapter underscores the functional role of job analysis in the assessment process, there is no doubt that the practice of job analysis requires a series of important methodological choices that can have both theoretical and practical implications. Thus, the second major section of this chapter focuses on the methodological aspects of job analysis. Its two subsections are centered on the sources and the methods of data collection, respectively.

ROLE OF JOB ANALYSIS IN THE ASSESSMENT PROCESS

This section reviews the job-analytic steps leading to the ultimate goal of developing assessment plans, beginning with the identification of job behaviors and their interplay with the work context in which these behaviors are carried out. Then, the process of drawing inferences concerning worker attributes on the basis of the aforementioned Job Behavior × Work Context interactions is discussed.

Identification of Job Behaviors

One of the first job-analytic approaches that focused on job tasks as the primary unit of analysis, functional job analysis was used as a basis for the now-defunct *Dictionary of Occupational Titles* (Fine & Cronshaw, 1999). Functional job analysis uses the basic structure of the English sentence to standardize the language of task descriptions. Thus, the basic structure of a task statement in functional job analysis includes the action verb, the object of the action, the source of information or instruction, and the results. However, the popularity of tasks as the unit of job analysis is also due to the task inventory approach, which was refined by Christal and his associates at the Air Force Human Resource Laboratory (Christal & Weissmuller, 1988).

Gael (1983) defined a task as "a unit of work performed by an individual that has a definite beginning and end and that results in a product or service" (p. 9). Task statements in the task inventory approach are usually worded using the elements of the English sentence as in Fine and Cronshaw's (1999) functional job analysis. The task inventory is formatted into a questionnaire (paper or computer-based) and distributed to a large sample of individuals, most often job incumbents and, in some cases, their direct supervisors, all of whom are asked to rate each task on certain scales (e.g., Sanchez & Levine, 1989). The scales might ask, for example, how often each task is performed and how critical it is. Of course, task inventories must follow the requirements for developing a valid survey measure (e.g., Miller, McIntire, & Lovler, 2011).

A task inventory provides a time-consuming but undoubtedly copious way to obtain information about a variety of work-related activities from numerous individuals in numerous work settings (Gael, 1983). The breadth of coverage provided by a task inventory is well suited to licensure and credentialing examinations, which test an individual's readiness for a wide variety of activities (Kane, 1982). Another benefit is that responses to a task inventory can be subject to statistical analyses to organize tasks into broader job components. Data from task inventories also provide empirical grounds for the development of assessment plans (Kane, 1997; Raymond, 1996), which can be decisive against legal challenges to the job relatedness of an assessment (Thompson & Thompson, 1982).

Customizing task statements to a particular job, however, results in an "apples and oranges" kind of issue when comparing jobs because a different set of task statements is used for every job. A more generic alternative that provides a common metric of work activities that cut across all occupations (Cunningham 1996) is provided by the 42 generalized work activities included in O*NET, the occupational network developed by the U.S. Department of Labor to replace the *Dictionary of Occupational Titles* (Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999). These 42 generalized work activities represent the synthesis of several classifications or taxonomies of work activity data (Cunningham & Ballentine, 1982; McCormick, Jeanneret, & Mecham, 1972). These generalized work activities, however, may understandably have less face validity than ad hoc task statements written specifically for the job in question and are often less acceptable to end users.

As mentioned, once tasks are spelled out in a task inventory, they are typically rated on importance, frequency, time spent, difficulty of learning, and other dimensions (Christal & Weissmuller, 1988; Sanchez & Fraser, 1992; Sanchez & Levine, 1989). However, the time-consuming and tedious process of rating a large number of tasks on multiple scales can adversely affect response rates and the validity of responses, so the question of how many tasks a task inventory should have is a pertinent one. The answer depends on the type of job under consideration, assuming that the inventory has been prepared to cover only one job. A simple job may be described with a dozen tasks, whereas more complex jobs may require hundreds of them. Because today's job boundaries evolve rapidly (Sanchez, 1994, 2000; Sanchez & Levine, 1999; Schneider & Konz, 1989; Siddique, 2004; Singh, 2008), it may be preferable to keep tasks at a relatively broad level of detail (Cunningham, 1996). However, although completing a long task inventory can lead to response distortion, task statements should not be so broad or ambiguous that they do not facilitate valid inferences regarding job requirements. For instance, the statement "handles customer complaints" fails to specify the degree of involvement of a customer service representative in this task because one is left wondering whether the agent simply records the complaint or tries to solve it.

Long inventories can induce respondent fatigue and lead to careless task ratings (Wilson, Harvey, & Macy, 1990). The degree to which this issue constitutes a threat to the validity of the data is sometimes assessed through the computation of veracity and carelessness indices. These indices include repeated items, items representing work activities that are known to be performed by all incumbents, and bogus items that are not part of the job at all (Pine, 1995). Green and Stutzman (1986) and Green and Veres (1990), however, found that different indices of carelessness do not always converge with each other. In addition, Dierdorff and Rubin (2007) found that these indices might not always capture carelessness but rather differences in incumbents' perceptions of their role. Therefore, one should be cautious about discarding respondents whose answers suggest carelessness according to these indices, which might unnecessarily reduce reliability and sample size. A possibly more direct approach to assess the validity of work activity data involves asking subject matter experts (SMEs) to estimate how well the inventory covers the scope of activities that make up the job. However, Wilson (1997) found that both incumbents and supervisors provided exceedingly confident judgments of inventory completeness, even when two thirds of the tasks had been removed. Perhaps the best way to combat carelessness is to motivate respondents to complete the inventory judiciously by impressing on them that their responses will lead to better selection of their coworkers. In addition, a meeting should be held with a panel of SMEs to review the preliminary task inventory before it is distributed for rating purposes. This panel should be asked to add tasks that are omitted in the inventory, eliminate obsolete or inaccurate tasks, and edit tasks as needed. The final goal is to reach an agreement among the experts regarding the final version of the inventory.

Given the aforementioned concerns about task inventory length, researchers have been motivated to study the overlap among different task scales. Prior research suggested that scales of criticality, overall importance, complexity or difficulty, and difficulty of learning load on the same factor, thus

providing relatively similar information on the relative importance of job behaviors (Friedman, 1990, 1991; Hubbard et al., 2000; Manson, Levine, & Brannick, 2000; Sanchez & Fraser, 1992; Sanchez & Levine, 1989). Task scales have largely used a relative format, which asks SMEs to compare tasks. McCormick (1960) first advocated this kind of relative format because he found that SMEs had difficulties using absolute scales that required them to allocate an exact portion of time spent on each work activity. However, Harvey (1991) argued against relative scales, which require ipsative judgments that preclude comparisons among jobs. Manson et al. (2000), nevertheless, conducted conventional construct validity tests of convergent and discriminant validity and concluded that relative and absolute scales provided virtually interchangeable information about job tasks.

Reliability of task inventory data is a necessary but not sufficient condition for validity. Studies of the reliability of task inventories have suggested that there is more agreement on data concerning molecular or specific job tasks than on data concerning broad job activities (Dierdorff & Wilson, 2003). In a similar fashion, Dierdorff and Morgeson (2009) noted higher interrater reliability estimates of tasks than of generalized work activities (i.e., .80 vs. .65). On the contrary, Voskuijl and van Sliedregt's (2002) meta-analysis yielded the opposite results, specifically that task ratings were less reliable than broader job behaviors (.62 vs. .29). Research comparing the reliability of decomposed (or task-based) and holistic (job-based) ratings may help clarify these conflicting findings, because it suggests that the higher interrater reliability of task ratings (Butler & Harvey, 1988; Gibson, Harvey, & Quintela, 2004; Harvey, Wilson, & Blunt, 1994) holds unless large numbers of probably tedious ratings of narrow job tasks are required (Cornelius & Lyness, 1980; Sanchez & Levine, 1994).

The reliability of task inventory data has been studied using both intrarater and interrater designs (Gael, 1983, p. 23). However, Dierdorff and Wilson's (2003) results challenged the assumption that the reliability of work activity ratings can be equivalently gauged through either interrater or intrarater designs. This finding is not altogether surprising, because Sanchez and Levine (2000) noted that interrater reliability estimates do not distinguish between variance accounted for by random sources and legitimate variance accounted for by the unique manner in which each incumbent approaches his or her job. Indeed, prior research has suggested that interrater differences may reflect not just perceptual differences but rather actual variations in the manner in which incumbents approach their job (Arvey, Davis, McGowen, & Dipboye, 1982; Arvey, Passino, & Lounsbury, 1977; Borman, Dorsey, & Ackerman, 1992; Dierdorff, Rubin, & Bachrach, 2012; Ford, Smith, Sego, & Quinones, 1993; Hazel, Madden, & Christal, 1964; Landy & Vasey, 1991; H. H. Meyer, 1959; Prien, Prien, & Wooten, 2003; Sanchez, Prager, Wilson, & Viswesvaran, 1998; Schmitt & Cohen, 1989; Silverman, Wexley, & Johnson, 1984; Tross & Maurer, 2000).

The notion of interrater disagreement as simply error is predicated on the hardly tenable assumption that jobs are stable across incumbents and over time (Cronshaw, 1998; Sanchez & Levine, 2009). An emerging stream of research has endorsed a more agentic view of the incumbent (Befort & Hattrup, 2003; Biddle, 1986; Dierdorff, Rubin, & Morgeson, 2009; Grant, 2007; Morrison, 1994; Roberts, Dutton, Spreitzer, Heaphy, & Quinn, 2005), who is seen as performing the job in accordance with his or her desired role identity, past experience, motivation, and goals. This agentic view is likely strengthened in today's organizations, in which electronic equipment has replaced humans in many standardized activities and in which empowering employees to perform tasks according to their own discretion is emphasized (Sanchez, 1994; Siddique, 2004; Singh, 2008). Wrzesniewski and Dutton (2001) named this process *job crafting,* which they defined as "the physical and cognitive changes individuals make in the task or relational boundaries of their work" (p. 179).

From an assessment point of view, understanding that interrater disagreement is a function not only of error but also of job individuation in the form of job crafting is important, because it enhances job analysts' ability to explain to end users the legitimate reasons behind observed disagreement among incumbents of the same job title, which

is known to hinder the face validity of the job analysis data (Jones et al., 2001; Sanchez & Levine, 2000). Specifically, such job individuation appears to be most widespread in data-oriented, complex jobs (Sanchez, Zamora, & Viswesvaran, 1997), jobs low in interdependence and routinization (Dierdorff & Morgeson, 2007), jobs in which equipment operation and direct contact are not involved, and jobs involving managerial activities (Lievens, Sanchez, Bartram, & Brown, 2010), probably because all of these factors provide plenty of opportunity for individuation in role enactment. From a practical perspective, when a less-than-desired level of interrater reliability is observed in the job analysis data, the analyst should consider the basis for disagreements. They may signal the need for training or prescriptions for changes in how incumbents should conduct their tasks (Sanchez et al., 1998).

Interaction Between Job Responsibilities and Work Context Demands

The very same responsibilities may call for different human attributes under different sets of working conditions. In other words, job responsibilities do not always call for the same human attributes to the same degree but sometimes interact with the context or situation in a manner that distracts, constrains, releases, or facilitates the expression of certain traits (Tett & Burnett, 2003). This way of thinking is consistent with interactional approaches (Mischel & Shoda, 1995, 1998), which explain within-person behavioral variability as a function of situation–response contingencies. This approach adds a third dimension to the two-way job Task × Worker attribute matrix that has dominated the making of inferences concerning attribute requirements in assessment applications of job analysis (Drauden & Peterson, 1974). Specifically, it suggests that job task–attribute relationships are moderated by contextual variables.

The concept of situational strength, which characterizes the extent to which situations restrict the expression of individual differences, is very relevant to the manner in which work context moderates job task–worker attribute relationships, especially in nonability domains such as personality (R. D. Meyer, Dalal, & Hermida, 2010; Mullins & Cummings, 1999; Weiss & Adler, 1984). R. D. Meyer, Dalal, and Bonaccio (2009) developed an O*NET-based measure of situational strength. Their meta-analysis of validity coefficients of personality measures supported the interactional hypothesis, because it yielded higher validity coefficients for occupations that had weaker situational strength scores.

The professional performance situation model is a general framework for conducting job analyses that incorporates some of the thinking inherent in interactional approaches. The professional performance situation model has been used primarily in the health professions (LaDuca, 1980, 1994; LaDuca, Engle, & Risley, 1978; LaDuca, Taylor, & Hill, 1984). It attempts to provide a comprehensive analysis of an occupation, including the major responsibilities, human attributes, and context in which practice occurs. The practice context includes social and technological factors, including the information required to perform job responsibilities. The professional performance situation model requires first the identification of the major responsibilities of the work role. Once these responsibilities are defined, the categories or conditions of performance within each responsibility are specified. For example, the same physician responsibility might be carried out quite differently depending on work context facets such as care setting (e.g., hospital vs. physician's office), organ system involved in the medical condition (e.g., endocrine, skeletal), and other possible contextual factors such as patient age, gender, and comorbidity and severity of the medical condition. The characteristic of the professional performance situation model that fits the interactional approach is that it aims to identify differences in the human attributes required to carry out a certain responsibility under various situations or conditions (LaDuca et al., 1984). Instead of relying on isolated tasks, the model specifies the entire scope of issues that an incumbent may need to solve as well as the situational factors that qualify which solutions are best for those issues. Another practical advantage is that the various cells of the model provide ad hoc information to develop assessments and performance exams that capture performance requirements across a variety of settings and conditions.

Clearly, this approach offers guides for the development of assessments that may be useful in the context of staffing and training.

The push for organizations to align their practices with their strategic goals has propelled the development of strategic competency modeling (Lucia & Lepsinger, 1999; Schippmann, 1999). Although the difference between job analysis and strategic competency modeling is still blurry (Schippmann et al., 2000), Sanchez and Levine (2009) argued that the primary purpose of competency modeling is to instill a shared understanding of the importance of certain strategic goals among employees so that they are motivated to perform their work assignment along strategic lines.

In spite of the rather vague early definition of *competency* as "any individual characteristic that can be measured or counted reliably and that can be shown to differentiate significantly between superior and average performers" (Spencer, McLelland, & Spencer, 1994, p. 4), several authors have argued for the view of competencies as broadly defined elements of the job performance space (Bartram, 2005; Lievens et al., 2010; Tett, Guterman, Bleier, & Murphy, 2000). Bartram's (2005) definition of competencies as "sets of behaviors that are instrumental in the delivery of desired results or outcomes" (p. 1187) represents this line of thinking. Lievens et al. (2010) also took the position that competencies are best classified as part of the performance space in their investigation of the sources of consensus in competency ratings.

Sanchez and Levine (2009) noted that most lists of competencies resemble broadly defined behavioral themes that represent what Becker, Huselid, and Ulrich (2001) termed *strategic performance drivers*. This view of competencies as behavioral themes that are instrumental in strategy implementation is consistent with the notion that competency modeling serves a very different purpose than job analysis (Sanchez & Levine, 2009). That is, borrowing the notion of volume from signal detection theory (Tett & Burnett, 2003), strategic competency modeling can be seen as an attempt to influence the work context, whose situational strength it tries to alter by raising the volume of specific channels or behavioral themes aligned with the organization's strategy. These loud signals should increase situational strength because they promote a shared understanding of the behavioral themes that are expected and rewarded (Bowen & Ostroff, 2004; Werbel & DeMarie, 2005). In this respect, Sanchez and Levine argued that strategic competency modeling is closest to a mechanism of social influence (Bowen & Ostroff, 2004; Chatman & Cha, 2003; O'Reily & Chatman, 1996; Werbel & DeMarie, 2005) than to traditional job analysis.

Sanchez and Levine (2009) suggested that strategic competency modeling research should focus on the extent to which competency models influence the workforce to promote day-to-day behavior along strategic lines, including the development of competency language that is accepted by end users, and the dissemination of behavioral examples that illustrate how employees can demonstrate strategic competencies in their jobs. Indeed, strategic aims and strategic forces within the organization are an underresearched facet of the work context that may indeed modify the type and the degree of human attributes required by the same job from organization to organization.

Drawing Inferences Concerning Worker Attributes

The job element method of job analysis first popularized the term *knowledge, skills, abilities, and other characteristics* (KSAOs), which refers to the human attributes necessary to perform a job (Primoff, 1975). *Knowledge* refers to an organized body of information, typically of a factual or procedural nature, applied directly to job performance. For instance, computer programmers need knowledge of specific languages. A *skill* refers to the competence to perform a learned, psychomotor act. An example is operating a forklift. An *ability* is a capacity to acquire the competence to perform an observable behavior or a behavior that results in an observable product. Firefighters, for example, are required to possess the physical ability to climb a ladder while carrying heavy objects. *Other characteristics* refer to personality factors, attitudes, and values needed to perform the job, such as being patient with an irate customer for a customer-contact job. KSAOs are rated on scales representing constructs such as how

important the element is in distinguishing the superior from the average employee, how much trouble is likely if the element is ignored when choosing among applicants, and to what extent the organization can fill its openings if the element is demanded. Structured questionnaires are available to help identify some KSAOs, such as personality (Raymark, Schmit, & Guion, 1997) and ability (Fleishman & Reilly, 1992) requirements.

From an assessment point of view, it is fair to say that abilities and other characteristics are the foundations on which knowledge and skills are developed. That is, knowledge and skills can be acquired through formal instruction and practice, whereas abilities and other characteristics are hard to modify through experience or time. Because abilities and other characteristics are not easy to acquire, work assessments need to ensure that job candidates possess the requisite ones for job training success. Standards such as the *Uniform Guidelines* (Equal Employment Opportunity Commission et al., 1978) warn against relying on easy-to-learn knowledge and skills when designing selection procedures, which should emphasize more difficult-to-learn abilities and other characteristics instead.

The derivation of KSAOs required for job performance is typically the last step in the process of assessment-oriented job analysis. To be job related, these human attributes are necessarily inferred by linking them to previously uncovered information regarding job behaviors (Hughes & Prien, 1989; Landy, 1988). Thus, SMEs identify the KSAOs required to perform each task. SME judgments can be dichotomous (yes–no) or can consist of ratings on a continuum that gauges their degree of relevance to each task. Although a two-way Task × KSAO matrix can be constructed to elicit these linkages, the number of judgments becomes overwhelming as the number of tasks and KSAOs increases. Landy (1988) suggested classifying tasks and KSAOs into broader categories to make this process manageable.

Developing KSAOs suitable to a job analysis questionnaire is not straightforward because KSAOs can be difficult to understand. The attribute-like nature of KSAOs is such that both questionnaire content and the choice of rating scales require careful consideration. Caution should be exercised to define KSAOs that are not too broad or otherwise ambiguous. For instance, *knowledge of labor law* in a job analysis of a human resource specialist may insufficiently convey the scope of knowledge required (does it include the most recent case law?) and the level required (consumer vs. interpreter of labor law?). KSAO statements should be carefully vetted for shared understanding of their meaning within a job. A helpful standard is to ensure that all KSAO specify three basic elements: (a) what type of KSAO it is (K, S, A, O), (b) in what context it is needed, and (c) at what level or degree of precision it is needed (Goldstein, Zedeck, & Schneider, 1993). SMEs charged with formulating KSAOs may ask themselves the following questions:

1. What are the attributes that distinguish the good from the bad performers?
2. What attributes explain why one worker is clearly superior to the other workers?
3. What are the underlying attributes that determined past examples of good and bad performance?

The scales used to evaluate KSAOs' importance and role in the assessment process should be similarly vetted. For example, the choice of scales should keep in mind that SMEs are often not neutral judges of the level and the extent to which certain KSAOs are required for their job. For instance, Raymond (2001) reported that the KSAO of advanced statistics, whose definition included examples, was rated moderately important or essential by 26% of those respondents who indicated that they never performed statistical tasks on the task inventory. Morgeson, Delaney-Klinger, Mayfield, Ferrara, and Campion (2004) tested the effects of impression management and self-presentation biases that Morgeson and Campion (1997) had previously proposed. Morgeson et al. concluded that ability statements were particularly prone to self-presentation biases because ability statements that were identical to task statements but were preceded by the phrase *ability to* drew higher ratings than their corresponding task statements.

Even though the use of rating scales with concrete behavioral anchors representing different

levels in the ability continuum has been advocated by Fleishman and his colleagues (Fleishman, Costanza, & Marshall-Mies, 1999; Fleishman & Reilly 1992) as a means of reducing subjectivity in ability ratings, Hubbard et al. (2000) noted that the behavioral anchors used in ability scales are problematic in that the level of the ability is potentially confounded with the degree of familiarity with the occupation to which the behavioral anchor pertains. That is the case with the anchor *reading professional surgery articles,* which may indeed be a relatively simple task for surgeons, and thus it should have never been placed at the top of the ability scale. Research in the performance appraisal domain (Tziner, Joanis, & Murphy, 2000) has also suggested that behavioral anchors sometimes interfere with rather than enhance ease of scale usage.

As mentioned, an implicit principle of the process through which KSAOs are inferred is that worker attributes should be rationally derived through job-related linkages with presumably more tangible features of the job such as job responsibilities and working conditions. The presence of job-unrelated sources of variance in worker attribute ratings is, therefore, considered undesirable. In this respect, Van Iddekinge, Putka, Raymark, and Eidson (2005) found considerable idiosyncratic variance in KSAO ratings. However, this finding is not altogether surprising because attribute inferences often capture unobservable constructs of a complex psychological nature, which require a large inferential leap that probably makes them particularly vulnerable to subjective influences. This conclusion is also supported by a study of O*NET ratings of personality requirements from 47,137 incumbents in more than 300 occupations sampled by the U.S. Department of Labor (Dierdorff & Morgeson, 2009), in which variance owing to raters was larger in personality traits (as much as 35%) than in job responsibility ratings (16%).

A more cumbersome procedure that resembles the one used in synthetic validity has also been used to identify appropriate KSAO measures (Arvey, Salas, & Gailluca, 1992; Goiffin & Woycheshin, 2006; McCormick et al., 1972; Sanchez & Fraser, 1994). Job component validation, which can be seen as a case of synthetic validity, involves statistically capturing the relationship between worker attributes and scores on job components, which typically represent job functions. The mechanical estimation of the worker attribute requirements of new jobs is made possible by applying the prediction equations that were calculated using prior data to the new job component data. For instance, LaPolice, Carter, and Johnson (2008) used O*NET data to estimate adult literacy requirements across occupations. Similarly, Jeanneret and Strong (2003) predicted general aptitude test scores using O*NET data. However, the measurement and statistical equivalence between scores determined through job component validation and those obtained through direct means should not be taken for granted (Harvey, 2011).

Developing Assessment Specifications and Plans

Once a job analysis has been completed, the results are used to develop a document referred to as *assessment specifications* or an *assessment plan.* The purpose of an assessment plan is to articulate the important standards that the assessment should meet, including a list of the job behaviors and KSAOs to be assessed. Most assessment plans also specify the type of measurement and the number of items allocated to each section. Assessment plans might also convey the difficulty and reading level of different sections of the assessment, the measurement method to be used for each section, and other features of the stimuli to be used in the assessment (Millman & Greene, 1989; Russell & Peterson, 1997). They may also specify the manner or order in which the section assessments may be administered and the way in which scores may be generated and combined.

Assessment plans serve multiple purposes. First, they provide direction to assessment developers, thereby ensuring continuity in assessment content and difficulty over time. Assessment plans also provide a framework for creating scale scores, equating test forms, and conducting other statistical analyses. In addition, they serve as evidence supporting the validity of inferences based on assessment scores. A test plan should also benefit from some degree of empirical support. For instance, a test plan should have higher interitem correlations within sections

than between sections and an acceptable pattern of convergent and discriminant validities among the scores representing the different sections. According to Raymond (2001), many shortcomings of online and computer-generated assessments can be attributed to the lack of specificity in their test plans.

A study by Nelson (1994) indicated that two states developed assessment plans whose content overlapped by only about 50%, in spite of the fact that both plans were developed on the basis of a job analysis of the same occupation, whose performance was unlikely to differ across state lines. In another study, Levine, Ash, and Bennett (1980) evaluated the outcomes of four very different methods of job analysis. In spite of considerable differences in cost, methodology, and other features, the test plans ultimately produced were remarkably similar. Levine et al. concluded that test plans seem to be heavily influenced by the insights and creativity of the person in charge of "leaping" from the job analysis data to the exams. The studies by Nelson and by Levine et al. illustrate that the linkages between the assessment plan and the tasks identified as being important through a job analysis can be sometimes tenuous. However, Manson (2004) found support for the practice of collecting at least moderately specific information such as the 10 most important tasks and 10 most important KSAOs, which indeed had an impact on the quality of the selection plans prepared on the basis of such job-analytic information.

The aforementioned studies reveal that even though job analysis is a critical part of the assessment development process, very little guidance is available concerning the rules for translating job analysis data into assessment plans (Kane, 1982, 1997; Raymond, 2001). Given that job analysis often serves as the primary evidence supporting the job relatedness of assessment results, guidelines and rules of evidence for the employment of job-analytic data in developing assessments plans are sorely needed. In the absence of such a body of rules, organizations lack the means to evaluate the extent to which their investment in job analysis has paid off by supporting truly job-related decisions. Ensuring the psychometric quality of job-analytic data appears insufficient because very similar information on job tasks and worker attributes may lead to very different selection plans when procedures for linking job-analytic information to assessment plans are left unspecified.

In contrast, rules that lay out the rationale for developing job-related assessment procedures, such as those formulated by Fine and Cronshaw (1999, pp. 133–136) regarding the use of task statements to develop employment interview questions and those proposed to link job tasks and KSAOs (Drauden & Peterson, 1974; Goldstein et al., 1993; Landy, 1988), may help to not only defend against legal challenges to what would otherwise be deemed arbitrary assessments but also heighten the probability that the assessment plan and measures derived from the analysis will be valid predictors of criteria.

Assessment plans can be based on job tasks or on KSAOs (Raymond, 2001). The first type is typically referred to as a *process-oriented assessment plan* because it is built around the various processes or job tasks involved in the job. An assessment plan organized around KSAOs is known as a *content-oriented plan* because it lists the various fields of knowledge and other human attributes to be covered in the assessment. To comply with job-relatedness provisions, the KSAOs included in content-oriented assessment plans are typically linked to job tasks following the procedures described in the section Drawing Inferences Concerning Worker Attributes earlier in this chapter (e.g., Landy, 1988).

Once the basic structure of an assessment plan has been articulated, the next step is to assign relative weights to each of its components (either job tasks or KSAOs). This assignment can be done through clinical judgment, that is, having SMEs judge the relative importance of each task or KSAO. However, relative weights can easily be computed by using the ratings gathered for each component, in which the relative importance of a given component is equal to the average rating for such a component divided by the sum of the ratings across all components. Because prior research has suggested that scales of criticality, importance, complexity, and difficulty are moderately intercorrelated to the point of loading on the same underlying factor (Friedman, 1990, 1991; Hubbard et al., 2000; Manson et al., 2000; Sanchez & Fraser, 1992), any combination of these scales thereof should provide relatively

interchangeable relative weights. Policy-capturing research, however, has suggested that most SMEs rely on a combination of criticality (defined as consequences of error) and difficulty of learning when judging overall task importance (Sanchez & Levine, 1989). However, researchers may use time-oriented scales such as time spent or frequency, which also tend to load on the same underlying factor, when they deem it necessary and legally sustainable to structure the relative weights according to time-oriented criteria. In addition, both criticality and time-oriented measures may be combined to form a composite index that captures both dimensions (Kane, Kingsbury, Colton, & Estes, 1989).

METHODOLOGICAL ASPECTS OF JOB ANALYSIS FOR ASSESSMENT PURPOSES

Assessments have profound consequences for the individuals who are being assessed. In the industrial and organizational psychology arena, such consequences often include drastic changes in employment conditions and compensation levels. It is therefore not surprising that the methodological choices underlying the assessment process are, more often than not, painstakingly scrutinized. This section reviews probably the two most critical methodological choices: who provides the job-analytic information and how the information is gathered.

Sources of Job-Analytic Information

Job incumbency has been the necessary and often sufficient criterion to select SMEs in traditional job analysis. However, the dynamic nature of today's work assignments calls for the inclusion of other sources of job analysis information (Bernardin, 1992). Different sources of information may be best qualified to provide information about distinct aspects of a job, in a manner akin to a 360-degree approach to job analysis. This approach can be cumbersome, because the recipients of each major job function should be first identified and then used as sources of information on solely those functions. For example, professional mystery shoppers may be best positioned to provide information on the customer service demands of jobs. Similarly, the difficulty of learning various tasks may be best judged by education specialists, and personality requirements such as tolerance for stress may be best assessed by those possessing a psychological background. However, alternate sources of work information should supplement rather than replace job incumbents, who have firsthand information about the job that is unavailable to others.

For example, Jones et al. (2001) found that job analysts made better predictions of worker attribute trainability than incumbents and students when trainability ratings were compared with actual changes in pre- and posttest learning measures. Similarly, Van Iddekinge, Raymark, and Edison (2011) correlated ratings of the extent to which KSAOs were needed at entry with ratings of perceived KSAO trainability made by a panel of psychologists. Their findings indicated that ratings of the more abstract ability and other characteristic attributes were less valid than those of more concrete knowledge and skill attributes. Taken together, these studies suggested that ratings of presumably more malleable knowledge and skills require different expertise than those of more stable abilities and other characteristics, as suggested by others (Harvey, 1991; Morgeson & Campion, 1997). Research has also shown that analysts produce more reliable job task-attribute linkage ratings than incumbents (Baranowski & Anderson, 2005).

Unlike items in other O*NET domains that are rated by incumbents, abilities and skills are rated by job analysts. This choice was justified by Peterson et al. (1999) using practical and theoretical arguments, such as the fact that sampling incumbents is expensive and that occupational analysts are better equipped to understand ability and skill requirements than are incumbents. O*NET researchers have found that incumbents provide higher ratings than analysts and that analysts' ratings are slightly more reliable than incumbents' ratings (Tsacoumis & Van Iddekinge, 2006). However, the O*NET rating materials handed out to analysts are screened to eliminate items that are thought to be irrelevant, such as knowledge, skills, education and training, and work styles, and these rating materials are also shortened in other ways. Even though these streamlined materials may simplify the information and thus increase rating reliability, such psychometric

gains may come at the expense of potentially relevant job information. In addition, skill and ability ratings in O*NET are formulated by occupational analysts without firsthand information acquired through incumbent interviews or incumbent observation. As suggested by Voskuijl and Sliedregt's (2002) meta-analysis, occupational analysts may produce more reliable ratings if such ratings are based on actual contact with job incumbents than if they are based on solely paper-based job descriptions as is the case with O*NET ability ratings. Other research has also suggested that increasing the amount of job information enhances the psychometric properties of job-analytic ratings (Harvey & Lozada-Larsen, 1988; Lievens, Sanchez, & De Corte, 2004).

Rater training has also been used to increase the validity of job-analytic ratings. Frame of reference has been the most widely used rater training format. This format attempts to standardize the raters' frame of reference. Lievens and Sanchez (2007) found that rater training increased interrater agreement and discriminant validity of competency ratings. Aguinis, Mazurkiewicz, and Heggestad (2009) found that frame-of-reference training reduced the correlation between SMEs' self-reported personality and job-analytic ratings of personality requirements and also lowered job-analytic ratings. Sanchez and Levine (1994) found that rater training aimed at reducing the potentially biasing influences of Tversky and Kahneman's (1974) representativeness and availability heuristics increased interrater agreement but only when the number of tasks rated was moderately small. Whether rater training interventions reduce idiosyncratic variance at the expense of important information regarding the manner in which incumbents experience job demands warrants further investigation.

Traditional data collection methods encourage the selection of SMEs with prior job experience. However, when jobs are new or are changing, incumbents lack direct job experience. The type of SMEs who may be best qualified to participate in the kind of future-oriented job analysis outlined by Schneider and Konz (1989) needs further research. For example, one can argue that prior job experience does not necessarily instill in incumbents a sense of how the job will change over time and that familiarity with emerging technologies may be helpful to formulate such judgments. Nevertheless, the validity of future-oriented ratings should be evaluated in longitudinal designs rather than taken for granted.

A related issue is the sample size required for reliable and valid job-analytic data. Not surprisingly, today's highly distributed and decentralized organizations resist convening large samples of SMEs to provide task or KSAO ratings. Sampling large numbers of SMEs is indeed time consuming and costly. Therefore, researchers have attempted to replace large samples with smaller SME panels. In two studies, Tannenbaum and Wesley (1993) compared ratings obtained from large samples of educators with those provided by a seven-SME panel. Even though the correlation between mean KSAO ratings provided by each source was high, intraclass correlations, which are sensitive to mean differences in ratings between two groups, were lower than .70. These results suggest that even though both groups produced similar rank orderings of KSAOs, they held somewhat different views regarding the absolute importance of each KSAO.

In a way, deciding on an adequate sample for job analysis should follow the same criteria as it would for any other survey. The ultimate goal is to draw a representative sample of the population of interest. Therefore, the sample may need to be expanded to include individuals whose work experiences are representative of the various components and conditions of the content measured in the questionnaire. Potential populations of interest may include practitioners, educators, managers, and experts in the field. Similarly, samples should be representative of the relevant population in terms of practice setting and demographic factors such as, at the very least, ethnic background, educational level, and gender. Diversity is particularly important because job incumbents' demographic characteristics may influence the type of issues and clients encountered while performing the job. In the field of credentialing, in which sample representativeness is critical, sample sizes commonly range from several hundred to several thousand individuals. Large samples improve the precision of statistical estimates by reducing sampling error, enhancing the generalizability

of results, and more generally lending credibility to the assessment results (Kane, Miller, Trine, Becker, & Carson, 1995). Obtaining a representative sample, however, is often fraught with practical obstacles. For example, many incumbents may be located in private practices or may not hold the correct job classification, which can increase the difficulty of constructing a comprehensive sampling frame. Similarly, obtaining adequate response rates can be challenging when lengthy questionnaires are used. Lack of motivation to participate in the analysis and protection of what may be viewed as proprietary information present additional hindrances. The use of total survey strategies for encouraging high response rates, including follow-ups and incentives, is recommended.

Methods of Collecting Information

Methods of data collection in job analysis include job observation, interviews, surveys, and examination of work records and documents, including prior job descriptions and occupational titles provided by O*NET and its predecessor, the *Dictionary of Occupational Titles*. It is also striking that in this digital era, traditional job analysis has not taken advantage of potentially rich sources of data such as the electronic performance monitoring systems that track mobile maintenance units, truck engines, and call center activity (e.g., number of calls handled and time spent on each call). However, the potential for invasion of privacy advises cautious usage of these technologies. For instance, Raymond (2001) noted how sensitive information from medical and insurance records can be valuable in analyzing health care occupations. In any case, there are legal reasons to document job analysis interviews, job observations, and other data collection activities in detail, such that an independent verification of what was done is feasible as required by the *Uniform Guidelines* (Equal Employment Opportunity Commission et al., 1978). Obviously, the source of information influences the method of data collection and even the resulting sample size.

Clearly, the type of job under analysis as well as the background of the job incumbents should be taken into account when selecting a method of data collection. For instance, a questionnaire may not be the best approach when analyzing jobs of a primarily physical nature, whose incumbents are not accustomed to surveys and paper forms. Similarly, job observation may not be the best methodology when analyzing the job of financial analyst, whose information-processing-oriented tasks are hardly observable.

Job analysis interviews should also be conducted in a manner that minimizes the likelihood of impression management and self-presentation biases summarized by Morgeson and Campion (1997). These interview practices begin with having interviewers introduce themselves in a nonintimidating manner, specify the unit that they represent, clarify the purposes of the interview while providing examples of applications of the information to be gathered (and possibly listing some of the purposes for which the information will not be used). Interviewers should note that they will interrupt when the activities being described seem unclear or out of sequence and also that the information will be subject to verification by other sources. To facilitate the job incumbent's recall of job tasks, the job analysis interviewer may proceed to ask job incumbents to describe a typical day at work, from beginning to end. Interviewers should consider interrupting job incumbents when the verb, object, or purpose of the tasks being described appear unclear, paying special attention to tasks that are described with ambiguous verbs implying a high degree of complexity such as *coordinate, direct, manage,* and *administer*. It is often easy to find a simpler verb that conveys the nature of the task in a more realistic manner.

A critical methodological issue in job analysis is the choice of a unit of analysis. Describing jobs at the task level is a common practice, probably because professional regulations such as the *Uniform Guidelines* (Equal Employment Opportunity Commission et al., 1978) call for the identification of important job behaviors. The need to develop a common metric that applies to every job, however, runs counter to the sole use of job tasks, which are too job specific and do not allow cross-job comparisons. To circumvent this problem, McCormick (1976) suggested describing jobs using generic worker-oriented items, which are very similar to the generalized work activities used in O*NET.

The rapid pace of today's businesses makes job descriptions obsolete very quickly, and therefore, broader units of analysis may be preferred. However, broad units may not always provide sufficient information to inform valid assessments while ensuring their defense against legal challenges.

SUMMARY AND CONCLUSIONS

In summary, job analysis has a long history of providing the information on job requirements that sets the foundation for work-related assessments. Job analysis has the potential to help organizations assemble a workforce that is most capable of achieving their business goals while complying with the job-relatedness provisions that emanate from laws and professional standards in the United States and other countries. In spite of ongoing business trends such as globalization and information technology that make work, work roles, and jobs much more flexible than they once were, job analysis still plays a critical role in ensuring that assessments respond to business necessities. Indeed, this review of research and practices concerning the manner in which job analysis informs assessment-related inferences suggests that job analysis can turn what would otherwise be guesswork into a directed, rational evaluation of requirements derived from the job behaviors and the context in which jobs are performed. Although the profession of industrial and organizational psychology has learned much about how to conduct job analysis and how to use it for developing valid assessments, much still remains to be researched on topics such as effective, evidence-based rules for translating job analysis findings into valid assessments.

References

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2009). Using Web-based frame-or-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology, 62*, 405–438. doi:10.1111/j.1744-6570.2009.01144.x

Americans With Disabilities Act of 1990, Pub. L. 101–336, 42 U.S.C. § 12101.

Arvey, R. D., Davis, G. A., McGowen, S. L., & Dipboye, R. L. (1982). Potential sources of bias on job analytic processes. *Academy of Management Journal, 25*, 618–629. doi:10.2307/256085

Arvey, R. D., Passino, E. M., & Lounsbury, J. W. (1977). Job analysis results as influenced by sex of incumbent and sex of analyst. *Journal of Applied Psychology, 62*, 411–416. doi:10.1037/0021-9010.62.4.411

Arvey, R. D., Salas, E., & Gialluca, K. A. (1992). Using task inventories to forecast skills and abilities. *Human Performance, 5*, 171–190. doi:10.1207/s15327043hup0503_1

Baranowski, L. E., & Anderson, L. E. (2005). Examining rating source variation in work behavior to KSA linkages. *Personnel Psychology, 58*, 1041–1054. doi:10.1111/j.1744-6570.2005.00234.x

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203. doi:10.1037/0021-9010.90.6.1185

Becker, B. E., Huselid, M. A., & Ulrich, D. (2001). *The HR scorecard: Linking people, strategy, and performance*. Boston, MA: Harvard Business School Press.

Befort, N., & Hattrup, K. (2003). Valuing task and contextual performance: Experience, job roles, and ratings of the importance of job behaviors. *Applied H.R.M. Research, 8,* 17–32.

Bernardin, H. J. (1992). An analytic framework for customer-based performance content development and appraisal. *Human Resource Management Review, 2*, 81–102. doi:10.1016/1053-4822(92)90019-M

Biddle, B. J. (1986). Recent developments in role theory. *Annual Review of Sociology, 12*, 67–92. doi:10.1146/annurev.so.12.080186.000435

Borman, W. C., Dorsey, D., & Ackerman, L. (1992). Time-spent responses as time allocation strategies: Relations with sales performance in a stockbroker sample. *Personnel Psychology, 45*, 763–777. doi:10.1111/j.1744-6570.1992.tb00967.x

Bowen, D. E., & Ostroff, C. (2004). Understanding HRM-firm performance linkages: The role of the "strength" of the HRM system. *Academy of Management Review, 29*, 203–221.

Brannick, M. T., Brannick, J. P., & Levine, E. L. (1992). Job analysis, personnel selection and the ADA. *Human Resource Management Review, 2*, 171–182. doi:10.1016/1053-4822(92)90010-N

Brannick, M. T., Levine, E. L., & Morgeson, F. P. (2007). *Job analysis: Methods, research, and applications for human resource management* (2nd ed.). Thousand Oaks, CA: Sage.

Butler, S. K., & Harvey, R. J. (1988). A comparison of holistic versus decomposed rating of Position Analysis Questionnaire work dimensions. *Personnel Psychology, 41*, 761–771. doi:10.1111/j.1744-6570.1988.tb00652.x

Chatman, J. A., & Cha, S. E. (2003). Leading by leveraging culture. *California Management Review, 45*, 20–34.

Christal, R. E., & Weissmuller, J. J. (1988). Job-task inventory analysis. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 2, pp. 1036–1050). New York, NY: Wiley.

Civil Rights Act of 1964, Pub. L. 88–352, 78 Stat. 241.

Cornelius, E. T., & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analysis by job incumbents. *Journal of Applied Psychology, 65*, 155–163. doi:10.1037/0021-9010.65.2.155

Cronshaw, S. F. (1998). Job analysis: Changing nature of work. *Canadian Psychology/Psychologie Canadienne, 39,* 5–13. doi:10.1037/h0086790

Cunningham, J. W. (1996). Generic job descriptors: A likely direction in occupational analysis. *Military Psychology, 8,* 247–262. doi:10.1207/s15327876mp0803_8

Cunningham, J. W., & Ballentine, R. D. (1982). *The General Work Inventory*. Raleigh, NC: Author.

Dierdorff, E. C., & Morgeson, F. P. (2007). Consensus in work role requirements: The influence of discrete occupational context on role expectations. *Journal of Applied Psychology, 92*, 1228–1241. doi:10.1037/0021-9010.92.5.1228

Dierdorff, E. C., & Morgeson, F. P. (2009). Effects of descriptor specificity and observability on incumbent work analysis ratings. *Personnel Psychology, 62*, 601–628. doi:10.1111/j.1744-6570.2009.01151.x

Dierdorff, E. C., & Rubin, R. S. (2007). Carelessness and discriminability in work role requirement judgments: Influences of role ambiguity and cognitive complexity. *Personnel Psychology, 60*, 597–625. doi:10.1111/j.1744-6570.2007.00085.x

Dierdorff, E. C., Rubin, R. S., & Bachrach, D. G. (2012). Role expectations as antecedents of citizenship and the moderating effect of work context. *Journal of Management, 38*, 573–598. doi:10.1177/0149206309359199

Dierdorff, E. C., Rubin, R. S., & Morgeson, F. P. (2009). The milieu of managerial work: An integrative framework linking work context to role requirements. *Journal of Applied Psychology, 94*, 972–988. doi:10.1037/a0015456

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88*, 635–646. doi:10.1037/0021-9010.88.4.635

Drauden, G. M., & Peterson, N. G. (1974). *A domain approach to job analysis.* St. Paul: Minnesota Department of Personnel, Test Research and Development Section.

Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38295–38309.

Fine, S. A., & Cronshaw, S. F. (1999). *Functional job analysis: A foundation for human resource management.* Mahwah, NJ: Erlbaum.

Fleishman, E. A., Costanza, D. P., & Marshall-Mies, J. (1999). Abilities. In N. G. Peterson, M. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 175–195) Washington, DC: American Psychological Association. doi:10.1037/10313-010

Fleishman, E. A., & Reilly, M. E. (1992). *Handbook of human abilities. Definitions, measurements, and job task requirements.* Palo Alto, CA: Consulting Psychologists Press.

Ford, J. K., Smith, E. M., Sego, D. J., & Quinones, M. A. (1993). Impact of task experience and individual factors on training-emphasis ratings. *Journal of Applied Psychology, 78*, 583–590. doi:10.1037/0021-9010.78.4.583

Friedman, L. (1990). Degree of redundancy between time, importance, and frequency task ratings. *Journal of Applied Psychology, 75*, 748–752. doi:10.1037/0021-9010.75.6.748

Friedman, L. (1991). Correction to Friedman (1990). *Journal of Applied Psychology, 76*, 366. doi:10.1037/0021-9010.76.3.366

Gael, S. (1983). *Job analysis: A guide to assessing work activities.* San Francisco, CA: Jossey-Bass.

Gatewood, R. D., Feild, H. S., & Barrick, M. (2008). *Human resource selection* (6th ed.). Mason, OH: Thomson/South-Western.

Gibson, S. G., Harvey, R. J., & Quintela, Y. (2004, April). *Holistic versus decomposed ratings of general dimensions of work activity.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Goffin, R. D., & Woycheshin, D. E. (2006). An empirical method of determining employee competencies/KSAOs from task-based job analysis. *Military Psychology, 18*, 121–130. doi:10.1207/s15327876mp1802_2

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis–content validity process. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 3–34). San Francisco, CA: Jossey-Bass.

Grant, A. M. (2007). Relational job design and the motivation to make a prosocial difference. *Academy of Management Review, 32*, 393–417. doi:10.5465/AMR.2007.24351328

Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology, 39*, 543–564. doi:10.1111/j.1744-6570.1986.tb00952.x

Green, S. B., & Veres, J. G. (1990). Evaluation of an index to detect inaccurate respondents to a task

analysis inventory. *Journal of Business and Psychology, 5*, 47–61. doi:10.1007/BF01013945

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 71–163). Palo Alto, CA: Consulting Psychologists Press.

Harvey, R. J. (2011, April). *Deriving synthetic validity models: Is R =. 80 large enough?* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago.

Harvey, R. J., & Lozada-Larsen, S. R. (1988). Influence of amount of job descriptive information on job analysis rating accuracy. *Journal of Applied Psychology, 73*, 457–461. doi:10.1037/0021-9010.73.3.457

Harvey, R. J., Wilson, M. A., & Blunt, J. H. (1994, April). *A comparison of rational/holistic versus empirical/decomposed methods of identifying and rating general work behaviors.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Nashville.

Hazel, J. T., Madden, J. M., & Christal, E. E. (1964). Agreement between worker–supervisor descriptions of the worker's job. *Journal of Industrial Psychology, 2*, 71–79.

Hubbard, M., McCloy, R., Campbell, J., Nottingham, J., Lewis, P., Rivkin, D., & Levine, J. (2000). *Revision of O*NET data collection procedures.* Raleigh, NC: National Center for O*NET Development. Retrieved from http://www.onetcenter.org/reports/Data_appnd.html

Hughes, G. L., & Prien, E. P. (1989). Evaluation of task and job skill linkage judgments used to develop test specifications. *Personnel Psychology, 42*, 283–292. doi:10.1111/j.1744-6570.1989.tb00658.x

Jeanneret, P. R., & Strong, M. H. (2003). Linking O*NET job analysis information to job requirement predictors: An O*NET application. *Personnel Psychology, 56*, 465–492. doi:10.1111/j.1744-6570.2003.tb00159.x

Jones, R. G., Sanchez, J. I., Parameswaran, G., Phelps, J., Shoptaugh, C., Williams, M., & White, S. (2001). Selection or training? A two-fold test of the validity of job-analytic ratings of trainability. *Journal of Business and Psychology, 15*, 363–389. doi:10.1023/A:1007804815480

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist, 37*, 911–918. doi:10.1037/0003-066X.37.8.911

Kane, M. T. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education, 10*, 5–18. doi:10.1207/s15324818ame1001_1

Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. (1989). Combining data on criticality and frequency in developing plans for licensure and certification examinations. *Journal of Educational Measurement, 26*, 17–27. doi:10.1111/j.1745-3984.1989.tb00315.x

Kane, M. T., Miller, T., Trine, M., Becker, C., & Carson, K. (1995). The precision of practice analysis results in the professions. *Evaluation and the Health Professions, 18*, 29–50. doi:10.1177/016327879501800103

LaDuca, A. (1980). The structure of competence in the health professions. *Evaluation and the Health Professions, 3*, 253–288. doi:10.1177/016327878000300302

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation and the Health Professions, 17*, 178–197. doi:10.1177/01632787940170020 4

LaDuca, A., Engle, J. D., & Risley, M. E. (1978). Progress toward development of a general model for competence definition in the health professions. *Journal of Allied Health, 7*, 149–156.

LaDuca, A., Taylor, D. D., & Hill, I. K. (1984). The design of a new physician licensure examination. *Evaluation and the Health Professions, 7*, 115–140. doi:10.1177/016327878400700201

Landy, F. J. (1988). Selection procedure development and usage. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 1, pp. 271–287). New York, NY: Wiley.

Landy, F. J., & Vasey, J. (1991). Job analysis: The composition of SME samples. *Personnel Psychology, 44*, 27–50. doi:10.1111/j.1744-6570.1991.tb00689.x

LaPolice, C. C., Carter, G. W., & Johnson, J. J. (2008). Linking O*NET descriptors to occupational literacy requirements using job component validation. *Personnel Psychology, 61*, 405–441. doi:10.1111/j.1744-6570.2008.00118.x

Levine, E. L., Ash, R. A., & Bennett, N. (1980). Exploratory comparative study of four job analysis methods. *Journal of Applied Psychology, 65*, 524–535. doi:10.1037/0021-9010.65.5.524

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology, 92*, 812–819. doi:10.1037/0021-9010.92.3.812

Lievens, F., Sanchez, J. I., Bartram, D., & Brown, A. (2010). Lack of consensus among competency ratings of the same occupation: Noise or substance? *Journal of Applied Psychology, 95*, 562–571. doi:10.1037/a0018035

Lievens, F., Sanchez, J. I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology, 57*, 881–904. doi:10.1111/j.1744-6570.2004.00009.x

Lucia, A., & Lepsinger, R. (1999). *The art and science of competency models: Pinpointing critical success factors in organizations.* San Francisco, CA: Jossey-Bass.

Manson, T. M. (2004). *Cursory versus comprehensive job analysis for personnel selection: A consequential validity analysis*. Unpublished doctoral dissertation, University of South Florida, Tampa.

Manson, T. M., Levine, E. L., & Brannick, M. T. (2000). The construct validity of task inventory ratings: A multitrait-multimethod analysis. *Human Performance, 13*, 1–22. doi:10.1207/S15327043HUP1301_1

McCormick, E. J. (1960). Effect of amount of job information required on reliability of incumbents' checklist reports. *USAF Wright Air Development Division Technical Note, 60–142,* 10.

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 651–696). Chicago, IL: Rand McNally.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 56*, 347–368. doi:10.1037/h0033099

Meyer, H. H. (1959). A comparison of foreman and general foreman conceptions of the foreman's job responsibilities. *Personnel Psychology, 12*, 445–452. doi:10.1111/j.1744-6570.1959.tb01336.x

Meyer, R. D., Dalal, R. S., & Bonaccio, S. (2009). A meta-analytic investigation into the moderating effects of situational strength on the conscientiousness-performance relationship. *Journal of Organizational Behavior, 30*, 1077–1102. doi:10.1002/job.602

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36*, 121–140. doi:10.1177/0149206309349309

Miller, L. A., McIntire, S. A., & Lovler, R. L. (2011). *Foundations of psychological testing: A practical approach* (3rd ed.). Thousand Oaks CA: Sage.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York, NY: Macmillan.

Mischel W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–268.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology, 49*, 229–258. doi:10.1146/annurev.psych.49.1.229

Mitchell, K. E., Alliger, G. M., & Morfopoulos, R. (1997). Toward an ADA-appropriate job analysis. *Human Resource Management Review, 7*, 5–26. doi:10.1016/S1053-4822(97)90003-6

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627–655. doi:10.1037/0021-9010.82.5.627

Morgeson, F. P., Delaney-Klinger, K., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks, and competencies. *Journal of Applied Psychology, 89*, 674–686. doi:10.1037/0021-9010.89.4.674

Morgeson, F. P., & Dierdorff, E. C. (2011). Work analysis: From technique to theory. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Vol. 2. Selecting and developing members for the organization* (pp. 3–41). Washington, DC: American Psychological Association.

Morrison, E. W. (1994). Role definitions and organizational citizenship behavior: The importance of the employee's perspective. *Academy of Management Journal, 37*, 1543–1567. doi:10.2307/256798

Mullins, J. M., & Cummings, L. L. (1999). Situational strength: A framework for understanding the role of individuals in initiating proactive strategic change. *Journal of Organizational Change Management, 12*, 462–479. doi:10.1108/09534819910300846

Münsterberg, H. (1913). *Psychology and industrial efficiency*. Boston, MA: Houghton Mifflin. doi:10.1037/10855-000

Nelson, D. (1994). Job analysis for licensure and certification exams: Science or politics. *Educational Measurement: Issues and Practice, 13,* 29–35.

O'Reily, C., & Chatman, J. (1996). Cultures as social control: Corporations, cults, and commitment. In L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 18, pp. 157–200). Greenwich, CT: JAI Press.

Pearlman, K., & Sanchez, J. I. (2010). Work analysis. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 73–98). New York, NY: Routledge.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association. doi:10.1037/10313-000

Pine, D. E. (1995). Assessing the validity of job ratings: An empirical study of false reporting in task inventories. *Public Personnel Management, 24*, 451–460.

Prien, K. O., Prien, E. P., & Wooten, W. (2003). Interrater reliability in job analysis: Differences in strategy and perspective. *Public Personnel Management, 32*, 125–141.

Primoff, E. S. (1975). *How to prepare and conduct job-element examinations* (U.S. Civil Service Commission Technical Study 75–1). Washington, DC: U.S. Government Printing Office.

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employee selection. *Personnel Psychology, 50*, 723–736. doi:10.1111/j.1744-6570.1997.tb00712.x

Raymond, M. R. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education, 9*, 237–256. doi:10.1207/s15324818ame0903_3

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14*, 369–415. doi:10.1207/S15324818AME1404_4

Roberts, L. M., Dutton, J. E., Spreitzer, G. M., Heaphy, E. D., & Quinn, R. E. (2005). Composing the reflected best-self portrait: Building pathways for becoming extraordinary in work organizations. *Academy of Management Review, 30*, 712–736. doi:10.5465/AMR.2005.18378874

Russell, T. L., & Peterson, N. G. (1997). The test plan. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 115–139). Palo Alto, CA: Davies-Black.

Sackett, P. R., & Laczo, R. M. (2003). Job and work analysis. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Comprehensive handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 21–37). New York, NY: Wiley.

Sanchez, J. I. (1994). From documentation to innovation: Reshaping job analysis to meet emerging business needs. *Human Resource Management Review, 4*, 51–74. doi:10.1016/1053-4822(94)90004-3

Sanchez, J. I. (2000). Adapting work analysis to a fast-paced and electronic business world. *International Journal of Selection and Assessment, 8*, 207–215. doi:10.1111/1468-2389.00150

Sanchez, J. I., & Fraser, S. L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology, 77*, 545–553. doi:10.1037/0021-9010.77.4.545

Sanchez, J. I., & Fraser, S. L. (1994). An empirical procedure to identify job duty–skill linkages in managerial jobs: A case example. *Journal of Business and Psychology, 8*, 309–325. doi:10.1007/BF02230375

Sanchez, J. I., & Levine, E. L. (1989). Determining important tasks within jobs: A policy-capturing approach. *Journal of Applied Psychology, 74*, 336–342. doi:10.1037/0021-9010.74.2.336

Sanchez, J. I., & Levine, E. L. (1994). The impact of raters' cognition on judgment accuracy: An extension to the job analysis domain. *Journal of Business and Psychology, 9*, 47–57. doi:10.1007/BF02230986

Sanchez, J. I., & Levine, E. L. (1999). Is job analysis dead, misunderstood, or both? New forms of work analysis and design. In A. Kraut & A. Korman (Eds.), *Evolving practices in human resource management* (pp. 43–68). San Francisco, CA: Jossey-Bass.

Sanchez, J. I., & Levine, E. L. (2000). Accuracy or consequential validity: Which is the better standard for job analysis data? *Journal of Organizational Behavior, 21*, 809–818. doi:10.1002/1099-1379(200011)21:7<809::AID-JOB28>3.0.CO;2-O

Sanchez, J. I., & Levine, E. L. (2001). The analysis of work in the 20th & 21st centuries. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. 1, pp. 71–89). Thousand Oaks, CA: Sage. doi:10.4135/9781848608320.n5

Sanchez, J. I., & Levine, E. L. (2009). What is (or should be) the difference between competency modeling and traditional job analysis? *Human Resource Management Review, 19*, 53–63. doi:10.1016/j.hrmr.2008.10.002

Sanchez, J. I., & Levine, E. L. (2012). The rise and fall of job analysis and the future of work analysis. *Annual Review of Psychology, 63*, 397–425.

Sanchez, J. I., Prager, I., Wilson, A., & Viswesvaran, C. (1998). Understanding within-job title variance in job-analytic ratings. *Journal of Business and Psychology, 12*, 407–419. doi:10.1023/A:1025046921171

Sanchez, J. I., Zamora, A., & Viswesvaran, C. (1997). Moderators of agreement between incumbent and non-incumbent ratings of job characteristics. *Journal of Occupational and Organizational Psychology, 70*, 209–218. doi:10.1111/j.2044-8325.1997.tb00644.x

Schippmann, J. S. (1999). *Strategic job modeling: Working at the core of integrated human resources.* Mahwah, NJ: Erlbaum.

Schippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., . . . Sanchez, J. I. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703–740. doi:10.1111/j.1744-6570.2000.tb00220.x

Schmitt, N., & Cohen, S. A. (1989). Internal analyses of task ratings by job incumbents. *Journal of Applied Psychology, 74*, 96–104. doi:10.1037/0021-9010.74.1.96

Schneider, B., & Konz, A. M. (1989). Strategic job analysis. *Human Resource Management, 28*, 51–63. doi:10.1002/hrm.3930280104

Siddique, C. M. (2004). Job analysis: A strategic human resource management practice. *International Journal of Human Resource Management, 15*, 219–244. doi:10.1080/0958519032000157438

Silverman, S. B., Wexley, K. N., & Johnson, J. C. (1984). The effects of age and job experience on employee responses to a structured job analysis questionnaire. *Public Personnel Management, 13*, 355–359.

Singh, P. (2008). Job analysis for a changing workplace. *Human Resource Management Review, 18*, 87–99. doi:10.1016/j.hrmr.2008.03.004

Smith, I. L., & Hambleton, R. K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice, 9*(4), 7–10. doi:10.1111/j.1745-3992.1990.tb00385.x

Spencer, L. M., McLelland, D. C., & Spencer, S. (1994). *Competency assessment methods: History and state of the art*. Boston, MA: Hay-McBer Research Press.

Stern, W. (1911). *Die Differentielle Psychologie in ihren methodischen Grundlagen* [Methodological foundations of differential psychology]. Leipzig, Germany: Barth.

Tannenbaum, R. J., & Wesley, S. (1993). Agreement between committee-based and field-based job analyses: A study in the context of licensure testing. *Journal of Applied Psychology, 78*, 975–980. doi:10.1037/0021-9010.78.6.975

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500–517. doi:10.1037/0021-9010.88.3.500

Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. J. (2000). Development and content validation of a "hyperdimensional" taxonomy of managerial competence. *Human Performance, 13*, 205–251. doi:10.1207/S15327043HUP1303_1

Thompson, D. E., & Thompson, T. A. (1982). Court standards for job analysis in test validation. *Personnel Psychology, 35*, 865–874. doi:10.1111/j.1744-6570.1982.tb02228.x

Tross, S. A., & Maurer, T. J. (2000). The relationship between SME job experience and job analysis ratings: Findings with and without statistical control. *Journal of Business and Psychology, 15*, 97–110. doi:10.1023/A:1007770919305

Tsacoumis, S., & Van Iddekinge, C. (2006). *A comparison of incumbent and analyst ratings of O*NET skills*. Retrieved from http://www.onetcenter.org/reports/SkillsComp.html

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. doi:10.1126/science.185.4157.1124

Tziner, A., Joanis, C., & Murphy, K. R. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perception, and ratee satisfaction. *Group and Organization Management, 25*, 175–190. doi:10.1177/1059601100252005

Van Iddekinge, C. H., Putka, D. J., Raymark, P. H., & Eidson, C. E. (2005). Modeling error variance in job specification ratings: The influence of rater, job, and organization-level factors. *Journal of Applied Psychology, 90*, 323–334. doi:10.1037/0021-9010.90.2.323

Van Iddekinge, C. H., Raymark, P. H., & Edison, C. E. (2011). An examination of the validity and incremental value of needed-at-entry ratings for a customer service job. *Applied Psychology, 60*, 24–45. doi:10.1111/j.1464-0597.2010.00425.x

Voskuijl, O. F., & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment, 18*, 52–62. doi:10.1027//1015-5759.18.1.52

Weiss, H. M., & Adler, S. (1984). Personality and organizational behavior. *Research in Organizational Behavior, 6*, 1–50.

Werbel, J. D., & DeMarie, S. M. (2005). Aligning strategic human resource management and person–environment fit. *Human Resource Management Review, 15*, 247–262. doi:10.1016/j.hrmr.2005.10.001

Wilson, M. A. (1997). The validity of task coverage ratings by incumbents and supervisors. *Journal of Business and Psychology, 12*, 85–95. doi:10.1023/A:1025066301244

Wilson, M. A., Harvey, R. J., & Macy, B. A. (1990). Repeating items to estimate the test–retest reliability of task inventory ratings. *Journal of Applied Psychology, 75*, 158–163. doi:10.1037/0021-9010.75.2.158

Wrzesniewski, A., & Dutton, J. E. (2001). Crafting a job: Revisioning employees as active crafters of their work. *Academy of Management Review, 26*, 179–201.

CHAPTER 24

THINKING AT WORK: INTELLIGENCE, CRITICAL THINKING, JOB KNOWLEDGE, AND REASONING

Nathan R. Kuncel and Adam S. Beatty

Measures of cognitive abilities have been an enduring mainstay for making hiring and trainability decisions in organizational and academic settings. The extensive literature on human cognitive abilities makes organizing a chapter a challenge. Because of the numerous high-quality chapters on human abilities in industrial and organizational psychology (e.g., Drasgow, 2002), this chapter includes a brief overview of well-established core findings followed by a more in-depth examination of newer questions and research directions for the future. A broad overview of the nature and structure, basic psychometric properties, and predictive power of scores obtained from ability measures and the linearity of relationships is presented. Finally, a discussion and critique of four different concepts in abilities—critical thinking, reasoning, working memory, and neuropsychological assessments—are presented.

PURPOSE OF ASSESSMENT

Cognitive ability measures are used for three major purposes in industrial and organizational psychology: selection, classification, and development or counseling purposes. For personnel selection, the measures are used to differentiate among those who are more able to do the job or more able to learn the tasks of the job. The main purpose of selection is to increase the performance of the selected group relative to the applicant group. Put another way, the goal is to hire the best-performing people. Classification involves attempting to address one or more organizational goals by assigning individuals from a group of recent hires to different jobs. One major goal is to increase the overall performance across the different jobs with attention to the relative importance of the jobs (send exceptional people to critical jobs). Other goals can be addressed with classification, including scarcity of acceptable performers for some jobs, applicant job preferences (with the goal of increasing worker satisfaction), and future staffing needs. The final purpose is to provide workers with feedback about their current capacities and to make recommendations for future training or development. These goals have influenced the types of measures that have been developed. Before discussing general models of intelligence and prominent measures used in the work setting, some key definitions are needed.

DEFINITIONS: *ABILITY, APTITUDE,* AND *ACHIEVEMENT*

A great deal of mischief can occur when there is no agreement about the definitions of ability, aptitude, and achievement. *Ability* is often used but with two very different implicit definitions. One definition is that general cognitive ability is an individual difference that merely reflects differences in current behavioral repertoire. Humphreys (1984) was in line with this definition, stating in his chapter on general cognitive ability that intelligence is "the entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature that [is] available at any one period of time" (p. 243). This definition is silent about the source of

DOI: 10.1037/14047-024
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

that behavior repertoire in terms of whether different levels of ability are developed or innate. A person's ability level therefore merely describes what the person can do. To the extent that a group has similar experiences and education, the tests scores will also reflect innate ability (but also environmental factors that affected development). An alternative definition of *ability* focuses on ability being innate capacity. Although innate capacity is scientifically interesting, all measures are measures of phenotype and are the result, to varying degrees, of innate ability, development, and environmental factors. Therefore, this chapter uses the definition of ability measures as being focused on current capabilities.

Achievement tests might be seen as ability tests used for the purpose of measuring recent change resulting from maturation, educational interventions, or training interventions. Aptitude measures are ability measures that are associated with gains on achievement measures, or at least with the prediction of future behavior. In practice, in industrial and organizational psychology, there is often little difference among ability, achievement, and aptitude measures. In many cases, most abilities simultaneously reflect some achievement over a given period of time (sometimes decades), and abilities, particularly broad ones, are associated with subsequent gains in knowledge or skill acquisition. Two organizing frameworks for thinking about these distinctions are worth reviewing.

Carroll (1993) presented a specific empirical definition for aptitude and achievement tests. Assume subjects are tested on two measures at Time 1. The first measure is an aptitude test and the second is an achievement test, the content of which will be the subject of subsequent training. After training, subjects are retested at Time 2 on both the aptitude test and the achievement test. In the purest case of an aptitude and an achievement test, Carroll (1993, p. 16) observed the following:

1. no reliable variance on the achievement measure at Time 1, that is, the subjects do not have any prior knowledge of the content to be learned;
2. reliable variance on the aptitude measure at Time 1, that is, one can measure real differences between subjects in aptitude;
3. no correlation between aptitude and achievement at Time 1, which is a result of all subjects having no knowledge of the content measured by the achievement measure;
4. no meaningful change in the aptitude measure at Time 2, that is, the training has no influence on the aptitude;
5. meaningful change in achievement scores from Time 1 to Time 2 and reliable variance in the achievement measures at Time 2, meaning subjects improve owing to training and improve on the achievement measures; however, because of Condition 1 there should be no correlation between Time 1 and Time 2 achievement measures; and
6. a significant correlation between aptitude at Time 1 and achievement at Time 2, that is, the aptitude measure predicts gains in achievement.

Although in practice relatively few situations fit this pure definition, they do occur (Carroll, 1974; Stanton, Koerth, & Seashore, 1930), primarily in the settings of foreign language acquisition and musical training. People know nearly nothing about the foreign language at Time 1, they learn during training, and other human abilities predict gains in foreign language knowledge and skill while remaining largely unaffected by the training. This framework is useful when considering the importance of content in personnel selection, discussed later.

A second complementary framework for considering the nature of ability measures was discussed by Lubinski and Dawes (1992). They argued for a widely held perspective that the authors consider to be the generally misleading distinction among ability, achievement, and aptitude measures and suggested considering measures using the following four characteristics (Cleary, Humphreys, Kendrick, & Wesman, 1975): (a) breadth of material sampled, (b) curriculum represented, (c) recency of learning sampled, and (d) purpose of the assessment.

In many cases, cognitive ability assessments in industrial and organizational psychology tend to have broad breadth and to sample historical curricula dating from primary and secondary school (e.g., math and language abilities developed in high school and earlier). This is especially the case for many

multiscale batteries. Measures of very specific job knowledge are narrower, have a focused curriculum, and reflect more recent learning. In many cases, the purpose of assessment is to make selection or trainability judgments, although some cognitive assessments are used for developmental purposes. It is worth noting that measures that fit Carroll's (1993) pure case of achievement would, almost by definition, be of narrow breadth, match the specific curriculum that is presented, and at Time 2 are measures of very recent learning. Their purpose, presumably, would be to quantify gains after training, education, or informal learning experiences. The combination of these frameworks could be used for hypothesis generation to further the understanding of how cognitive abilities function and develop in work settings.

FACTOR STRUCTURE

Modern factor models of human abilities arrive at a hierarchical solution with a general factor on top with one or more layers of increasing numerous and specific factors below. Until recently, the dominant theory was the three-stratum model proposed by Carroll (1993) that was, in part, based on the theory of fluid and crystalized intelligence (originally proposed by Cattell, 1941, 1971). On the basis of an exhaustive review and reanalysis of more than 450 datasets of human abilities, Carroll concluded that abilities could be reasonably organized into a second stratum consisting of fluid, crystallized, general knowledge, visual–spatial ability, short-term memory, long-term memory, cognitive processing speed, and decision speed. Below these were typically two or more primary mental abilities, and above the second stratum was general mental ability (GMA).

Theories anchored in the theory of fluid and crystallized intelligence have generally fit both factor-analytic data as well as most research within a broader nomological network (see Chapter 4, this volume, for an elaboration of a nomological network). Fluid intelligence was argued to be invested by individual into acquiring crystallized intelligence. Crystallized intelligence includes traditional school learning but also applies to job knowledge (vocational knowledge) as well as knowledge and skill acquired for hobbies or other outside interests (avocational knowledge). Anyone who has met someone with an overwhelmingly intense interest in model railroading, baking sourdough bread, or Star Wars lore has experienced a person with focused crystallized intelligence or avocational knowledge.

The concept of investment theory has been an enduring theoretical concept in human abilities because of field, longitudinal, and laboratory research that has demonstrated the importance of fluid intelligence in acquiring knowledge and skill over time (Ackerman, 1987; Kuncel & Hezlett, 2007a, 2007b). Additionally, research on the influence of aging on human cognitive abilities has found larger declines for abilities conceptually linked to fluid intelligence and smaller or zero decline for abilities conceptually linked to crystallized intelligence.

Johnson and Bouchard (2005) have proposed a conceptual and empirical challenge to the three-stratum theory they named VPR, made up of verbal, perceptual, and image rotation abilities. It too is organized hierarchically, with VPR falling under the highest stratum, GMA or *g*. At the second stratum are narrower measures falling under VPR, and below these are individual tests. However, any given measure may be a mixture of second-stratum abilities that would be identified by cross-loadings in a factor analysis. Johnson and Bouchard's proposal is particularly compelling for three reasons that should be adopted by other scholars. First, they examined sizable datasets with a diverse set of ability measures. Second, their study pitted different models against each other. Rather than simply putting forth a model or comparing the preferred model with a straw man model that decades of research would reject outright (e.g., one factor with no subfactors), they compared the relative fit of different competing models. Third, they reviewed evidence from the broader literature and fit their model into the nomological network. Specifically, the *g*–VPR distinctions fit heritability and genetic evidence (Johnson et al., 2007) and neuroscientific evidence for separate brain modules for processing language and perceptual information.

NOTABLE MEASURES

Development of measures for work-related purposes have typically been oriented to knowledge, skill, and

ability requirements of jobs rather than being based on specific theories of intelligence. Because of the interest in what people can do as well as in their ability to acquire job knowledge and skill on the job, one can argued that test development of cognitive ability measures in work settings has followed the theory of crystallized and fluid intelligence.

Several measures have been heavily used in practice and for research. These measures are all group-administered tests that are needed to efficiently evaluate large numbers of people. The General Aptitude Test Battery (McCloy, Russell, & Wise, 1996) from the U.S. Department of Labor was first published in 1947. This form has been used for hundreds of studies (e.g., Bemis, 1968) and had nine aptitudes: general learning ability, verbal aptitude, numerical aptitude, spatial aptitude, form perception, clerical perception, motor coordination, finger dexterity, and manual dexterity.

Similarly, the Armed Services Vocational Aptitude Battery (Defense Manpower Data Center, 2006) shares a similar structure, consisting of a combination of verbal measures (word knowledge, paragraph comprehension), quantitative measures (mathematics knowledge, arithmetic reasoning), and several measures designed to evaluate more specific domains (general science, electronics information, auto and shop information, mechanical comprehension, and object assembly).

Both of these examples are consistent with overall themes in cognitive assessments noted by Hunt (2011), who suggested that batteries of tests often contain scales on language use, visual–spatial reasoning, mathematical reasoning, and deductive and inductive reasoning. In occupational settings, a theme of scales developed to contain more job-specific abilities can be added. The importance of these occupationally specific scales and even content as a whole has been a topic of ongoing interest and debate in work settings.

CONTENT- AND DOMAIN-SPECIFIC KNOWLEDGE

In work settings, a long history of research has examined the importance of two related topics: differential weighting schemes or the degree to which specific ability variance can increment GMA in predicting training and work performance. A persistent finding has been that in multitest batteries, differential weighting of tests has little practical importance in predicting either job performance or training outcomes. This pattern of findings has been interpreted as demonstrating the general unimportance of test content for applied use (e.g., Murphy, Dzieweczynski, & Yang, 2009).

The applied exceptions to this pattern are rare and, generally, narrowly focused. A body of research in work settings has suggested modest predictive gains for spatial abilities and mechanical abilities (e.g., Muchinsky, 1993) for jobs requiring such abilities. Zeidner and Johnson (1994) found that specific abilities substantially improved classification efficiency. However, increments are typically very modest and would, in practice, have little effect on the people hired in a top–down selection. Outside of work settings, specific knowledge is an especially good predictor of subsequent performance in training or educational programs. For predicting success in foreign language training, language aptitude increments general cognitive ability (Silva & White, 1993). Kuncel, Hezlett, and Ones (2001) examined the predictive power of GRE Subject exams and found that they were consistently superior predictors across several measures of performance. The most noteworthy prediction was for the criterion of degree attainment for which prior grades and GRE General exams were very modest predictors. In contrast, the GRE Subject exams were, comparatively, excellent predictors of finishing graduate school. These results may reflect indirect interest measurement through an assessment of subject area knowledge as well as being evidence of effective prior learning.

The general finding that content and weighting (e.g., Schmidt, 2002) are not terribly important is valuable in providing practical guidance for existing measures and batteries. However, a few aspects of these studies are worth noting. First, the field has examined multitest batteries that contain specific ability measures that were selected to be broadly relevant to a range of jobs. That is, the comparisons are not based on randomly selected sets of abilities from the vast array that have been identified. Instead, the

contest is between a set of good candidates. Second, and most important, this body of research has tacitly assumed that the applicant group has no subpopulations within it. That is, the assumption is that there are no groups who effectively lack the ability to perform effectively on one measure while performing effectively on the other.

We propose that the presence of the following ability and situational characteristics would tend to increase the incremental predictive power of an ability measure over general mental ability for predicting work performance or training success:

1. The ability measure captures critical job knowledge and skill. Performance on the job or in training is particularly reliant on the content domain sampled by the ability measure.
2. Exposure and opportunity to learn from formal instruction are not equal across all job applicants. As exposure heterogeneity increases, incremental predictive validity should increase. That is, not everyone has been taught the knowledge or skill captured by the measure.
3. Opportunities to learn the requisite knowledge or skill are rare in typical environments. Facility in interpreting literary texts could be acquired from many sources including school, book groups, or free public lectures. Nuclear submarine technician skills could not be acquired in this way.
4. Consistent with Characteristics 2 and 3, after hiring, employees are not trained to a performance criterion on the job-relevant knowledge and skill (e.g., military occupations).
5. Compensatory knowledge cannot be applied to permit performance. That is, there is only one way to do the job correctly, and other knowledge will not get one around this fact.
6. The ability measure is associated with a large interest component in the subject matter that would tend to yield better volitional choices.
7. The ability measure captures knowledge that is necessary to permit additional learning after hiring. For example, if learning advanced statistical techniques is needed after hiring, those without a good working knowledge of algebra will be in trouble on Day 1.

We hypothesize that such situations are rare but worth investigating. This framework may be of importance for understanding prediction bias as outlined in Kuncel and Klieger (2012). It seems likely that such situations will typically involve job-specific knowledge. As previously noted, multitest batteries for personnel selection are logically focused on a subset of human abilities that are of particular interest in work settings. Existing batteries also sample domains that are a part of the educational background of most, if not all, job applicants.

In addition, societies have generally shielded themselves from the worst-case scenarios through licensing and credentialing. For example, would it make any sense to hire doctors or nurses based on a GMA test without first ensuring that they are licensed? Licensure acts as a check on shared learning experiences. Grossbach and Kuncel (2009) reported an uncorrected correlation of .46 between SAT scores and subsequent performance on the National Council Licensure Examination required for licensure as a registered nurse. It contains highly job-specific items similar to this one:

> Which medications are associated with a higher risk of bleeding? Note that more than one may apply.
>
> A. Lovenox
> B. Enalapril
> C. Warfarin
> D. Aspirin
> E. Lansoprazole

Within a sample of nursing students, one would expect to find that ability tests and job knowledge tests will produce similar rank ordering of nurses and result in relatively little difference in the subsequent performance of a selected group. However, all of the nurses have been trained at accredited programs, have covered highly uniform curricula, and have been required to pass a standardized licensing examination. If one takes the "content does not matter" argument seriously, then SAT scores could reasonably be substituted for a knowledge measure without attention to licensure. For example, many people earn respectable SAT scores but have little medical knowledge. In the absence of licensure, inattention to content might result in hiring a

number of clever nurses, accountants, and industrial and organizational psychologists. The consequences could be dire.

A simulation of a related scenario is presented next, based on real data from a study in the literature. These data illustrate the case that content is important when groups differ in access to the necessary knowledge and skill. Grigerenko et al. (2004) examined the abilities of high-school-age children in two remote Alaskan areas. The first group, called the *semiurban group,* lived in a small town under conditions more similar to the majority of the United States. The second group, called the *rural group,* lived much more traditionally, continuing to hunt and collect food from the land. Both groups attended school. The study collected traditional cognitive ability measures on both groups as well as a measure of land, sea, and river knowledge that reflected very specific abilities for hunting and gathering. This measure included questions about where to fish, what to gather, and how to find game. Finally, adults were independently asked to rate the young adults on hunting performance. The semiurban students, on average, scored higher on the traditional ability measures and much lower and more variably on the land, sea, and river knowledge measure. Both measures remained correlated, but weakly, reflecting a lack of investment in acquiring this skill for most of the individuals studied. In contrast, among the rural students, knowledge scores were much higher and were more strongly correlated with traditional cognitive ability measures, reflecting the dual investment in both schooling and traditional skills.

To test the notion that content does not matter among correlated tests, we used the means, variances, and correlations to simulate equal-sized applicant groups for the job of hunter, which is a job of importance if one wishes to eat well in the rural setting. Applicants were selected using either the GMA test or the knowledge measure. Across 1,000 replications, if the top half (signed rank = .50) were selected using the GMA test, the average performance declined in the selected group ($Z = -.08$). In contrast, use of the knowledge measure resulted in improved average performance ($Z = .08$). Scatterplots for some of the simulated data display the shifts in the groups depending on which predictor is used (see Figure 24.1).

Clearly, content does matter. Previous research did not consider scenarios in which there are, effectively, different populations. The work by Murphy et al. (2009) assumed no subpopulations with unequal means, variances, and correlations. Their assumption is often true with the types of measures typically used in hiring situations. However, it should not be taken to be a universal truth as illustrated and discussed here. Similar effects as in the extreme hunting example may be observed for other knowledge domains that are not part of typical formal education. Murphy et al. (2009) considered the case of when selection measures tap knowledge and skill that result from shared educational backgrounds. The contradictory results seen here reflect an unusual but psychologically important scenario.

CRITERION-RELATED EVIDENCE OF VALIDITY

The relationship between measures of cognitive ability and subsequent outcomes has been a topic of intense study for more than 100 years. A substantial body of evidence has accumulated for the relationship between cognitive ability and work performance, formal training outcomes, and academic achievement at all levels.

Work Performance

Cognitive ability measures are consistently some of the strongest predictors of subsequent work performance (Kuncel & Hezlett, 2010; Schmidt & Hunter, 1998). Three major explanations have been offered for this finding. The first is that as measures of current capabilities, all else equal, those higher in general cognitive ability simply know more and are able to do more, leading to better performance. The second explanation is that GMA is associated with the capacity to learn effectively; therefore, those higher in GMA will be superior at acquiring job-specific knowledge, leading to superior performance. Finally, the third explanation notes that GMA is associated with the ability to rapidly and accurately process information and that this ability leads to superior performance. Given the intertwined nature

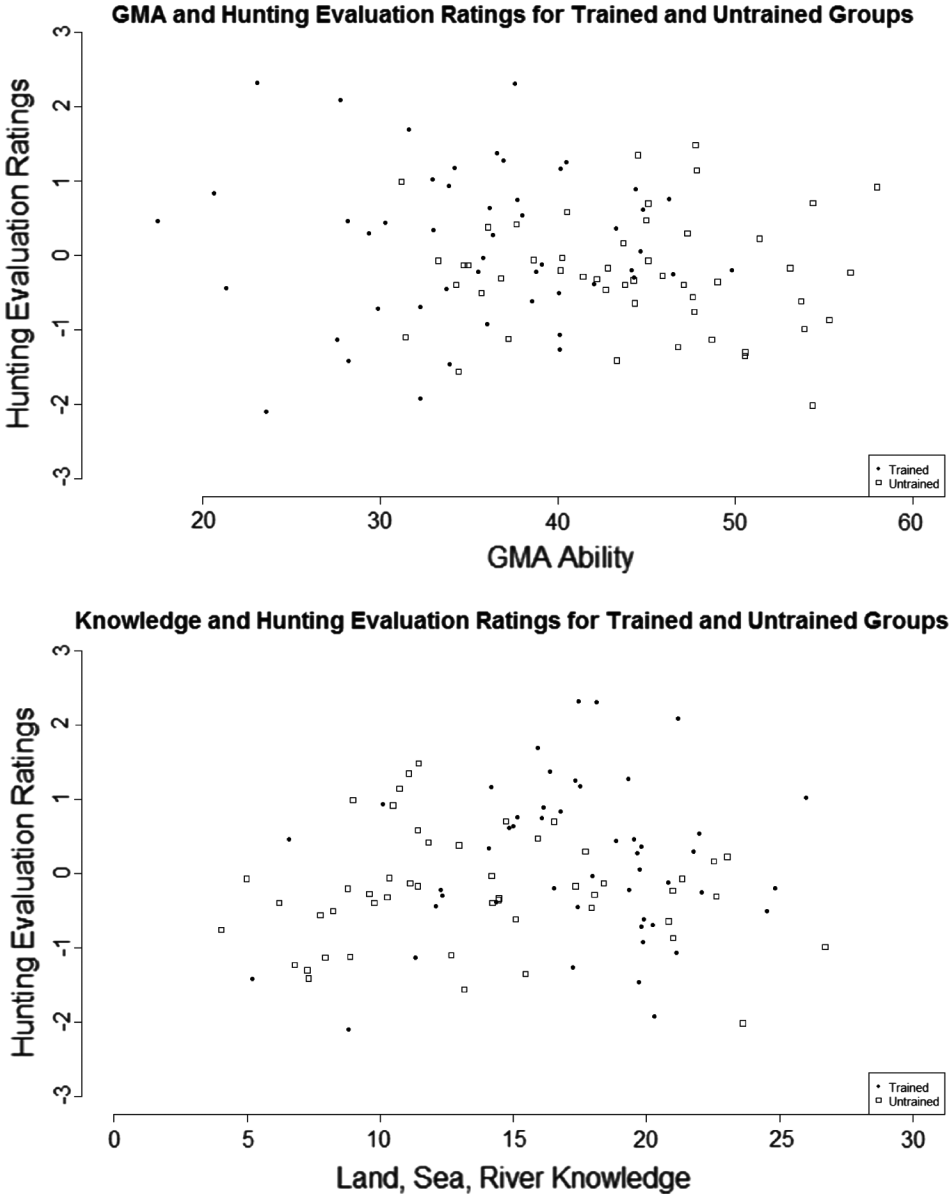

FIGURE 24.1. Simulated evaluation ratings for trained and untrained groups based on data from Grigorenko et al. (2004) examining abilities of high-school-age children in two remote Alaskan areas. GMA = general mental ability.

of human abilities, it should come as no surprise that evidence exists to support all of these explanations. Figure 24.2 summarizes results from multiple meta-analytic sources on the predictive power of general cognitive ability for work outcomes.

The pattern of relationships is not constrained to the United States and appears to hold across nations as well (e.g., Salgado et al., 2003). Research has indicated that the correlation between GMA measures extends beyond task performance to other aspects of performance, including objective leadership behavior (Judge, Colbert, & Ilies, 2004) and evaluations of creativity (Kuncel et al., 2004). Although by no means the only determinant of work performance, cognitive ability has remained an important predictor across decades of research. The predictive power of cognitive ability for work performance is strongest for the highest complexity jobs and lower for less complex jobs (Ones, Viswesvaran, & Dilchert, 2005; Salgado et al., 2003). Many other measures improve prediction of job performance when added to cognitive ability (Schmidt & Hunter, 1998).

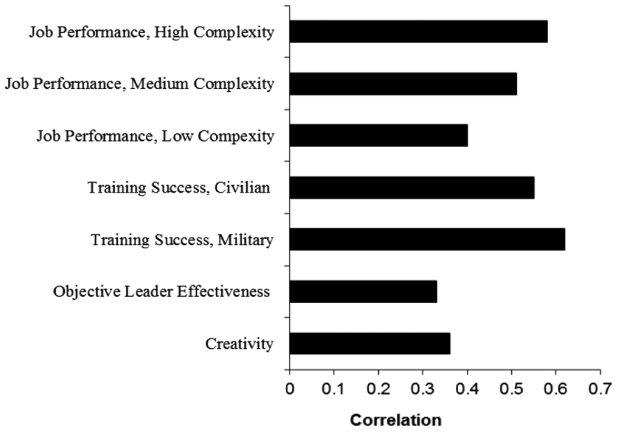

FIGURE 24.2. Meta-analytic estimates of the relationship between general cognitive ability and work and training performance.

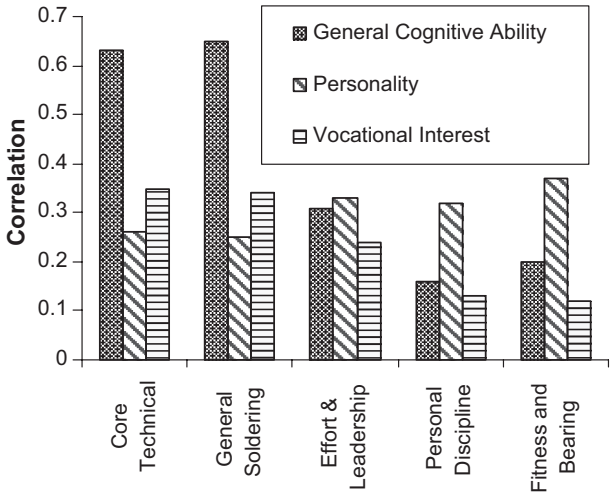

FIGURE 24.3. U.S. Army data performance prediction, Project A. Data from McHenry et al. (1990).

Often, the issue of other predictors is raised with the question, "What is the most important human characteristic?" The best answer is, "It depends." Results from the large selection and classification project of the U.S. Army nicely illustrates this fact in Figure 24.3. McHenry, Hough, Toquam, Hanson, and Ashworth (1990) compiled data across a large sample with multiple measures of different aspects of job performance. For some aspects of performance, the single best predictor was cognitive ability, whereas for others it was vocational interests or personality. A complete understanding of job performance requires understanding the interplay among multiple individual differences (and situational factors as well).

Academic Performance

Grades are well predicted by ability test scores at all levels. Research in primary and secondary schools has long established correlations between .40 and .50 (Matarazzo, 1972; Neisser et al., 1996), and these estimates are based on thousands of data points. General cognitive ability is also associated with graduating from high school as well. At the college level, test scores are similarly correlated with college grades. One study with more than 155,000 subjects yielded an operational validity of .47 for college grades (.53 for the full test-taking population assuming there was no self-selection into colleges and universities; Sackett, Kuncel, Arneson, Cooper, & Waters, 2009). In college, although test scores are not as strong a predictor as high school grades, they provide substantial incremental predictive power over high school grades alone (e.g., Berry & Sackett, 2009).

The association between cognitive ability and retention, graduation from college, or both depends on the research question asked. One line of research has concluded that a positive but trivially large relationship exits, particularly after controlling for other variables (Bowen, Chingos, & McPherson, 2009). A second line has indicated an important relationship between the two (Mattern & Patterson, 2009), in which 95.5% of high-scoring students return for the second year, whereas only 63.8% of low-scoring students return. This difference in findings is explainable by the nature of the analysis. Bowen et al. (2009) examined graduation at colleges and universities, and Mattern and Patterson (2009) examined retention across schools. Rephrased, the data suggested that high-scoring students tend to go to institutions from which more students graduate, but within any given institution, the relationship is very small. Whether this pattern can be attributed to individual, peer group, or school effects remains to be definitively determined.

The relationship between ability tests and performance in graduate and professional schools was reviewed by Kuncel and Hezlett (2007a, 2007b) in a study with more than 600,000 students and thousands of individual studies across all major graduate admissions tests (GRE, Pharmacy College

Admission Test, Graduate Management Admission Test, Miller Analogies Test, Law School Admission Test, Medical College Admission Test). This review included results for eight measures of academic performance including grades, faculty evaluation of performance, research productivity, and licensing examinations. For all ability measures and all performance measures, correlations were positive and often substantial. More motivationally determined indicators of performance such as degree attainment were less well predicted, whereas faculty judgments of student performance, grades, and comprehensive and licensing examinations were well predicted. Unlike the collegiate setting, graduate school outcomes were typically better predicted by ability test scores than by college GPAs. Results from the GRE Subject tests indicated that they were the single best predictor for all outcomes (for those fields that have available and use the GRE Subject tests). Figure 24.4 displays results across multiple studies for both college and graduate school performance measures.

Training Performance

Given the relationships between ability measures and learning in kindergarten through 12th grade, college, and graduate school, and the extensive empirical literature demonstrating the relationship between cognitive ability and complex skill acquisition in laboratory settings (e.g., Ackerman, 1987), it is unsurprising that performance in formal training programs is predicted by cognitive ability measures. Among several characteristics related to training success, cognitive ability is consistently a strong predictor (Campbell & Kuncel, 2001). Colquitt, LePine, and Noe (2000) presented a model that combines self-efficacy, cognitive ability, personality, and job and situation factors in explaining training and subsequent outcomes. Cognitive ability was associated with pretraining self-efficacy, acquisition of knowledge and skill, and posttraining self-efficacy.

LINEARITY

Although positive correlations are consistently observed between measures of ability and subsequent performance, the shape of the relationship has remained an ongoing topic of speculation. Three general relationships have been proposed. The first is a simple linear relationship in which

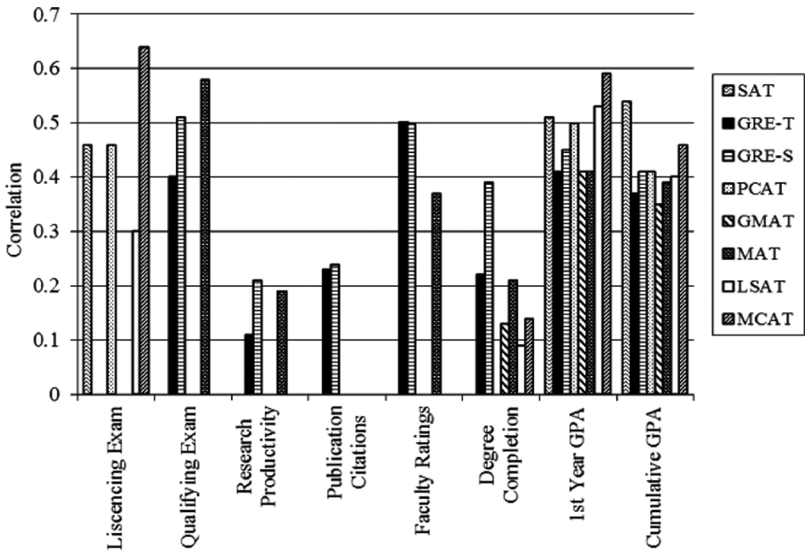

FIGURE 24.4. Meta-analytic estimates of the correlation between cognitive ability measures and academic performance. GRE-T = GRE General test; GRE-S = GRE Subject test; PCAT = Pharmacy College Admission Test; GMAT = Graduate Management Admission Test; MAT = Miller Analogies Test; LSAT = Law School Admission Test; MCAT = Medical College Admission Test.

increases in cognitive ability are associated with consistent gains in performance. More is better. The second is a plateau or asymptotic relationship in which the importance of ability holds, but only up to a certain point, also called *minimum competence relationship*. After some minimum level, the relationship flattens partially or completely. The final relationship is one in which ability beyond a certain point is said to actually become a liability. This is a "too much of a good thing" relationship. Conventional wisdom has generally endorsed the "good enough" asymptotic relationship. Ironically, the relationship does appear to be nonlinear but in the opposite direction. If anything, the strength of the relationship increases as cognitive ability increases. Research by Arneson, Sackett, and Beatty (2011) across large sample educational and work data sets found no evidence for a plateau and, if anything, an acceleration in the lines between ability and subsequent performance. The acceleration of the lines at the higher end of the ability continuum may be due to the findings that ability is more strongly correlated with performance in higher complexity work (Ones et al., 2005; Salgado et al., 2003) and academic settings (Shen et al., 2012).

This new evidence is consistent with previous research in both work and academic settings. Coward and Sackett (1990) examined a series of 174 samples containing more than 36,000 workers to test for nonlinear relationships between an ability test (General Aptitude Test Battery) and subsequent evaluations of job performance. The results indicated roughly chance-level incidents of significant nonlinear relationships. Although the data were constrained to less than 2.5 standard deviations and power to detect nonlinear relationships was not high, even with the samples used by Coward and Sackett (1990), the evidence argues strongly against any sizable nonlinear effects.

More recently, an extreme test of the hypothesis presented by Lubinski (2009) on the performance of people who test in the top 1% of cognitive ability arrived at the striking finding that those in the top quarter of the top 1% had substantially better work and educational outcomes (patents, publications, earning a doctorate, income) than those who were in the bottom quarter of the top 1% (see Figure 24.5). Even within a very narrow and extreme band of ability performance and success, differences were observed.

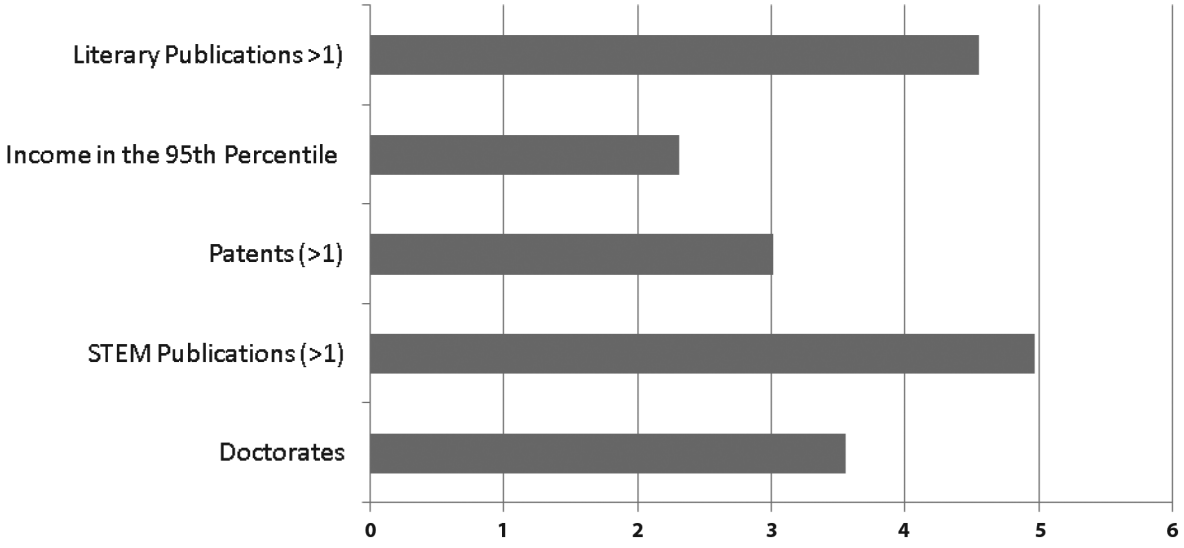

FIGURE 24.5. Odds ratios comparing accomplishments of the top and bottom quartile within the top 1% in cognitive ability 25+ years after identified at age 13. Data from Lubinski (2009). STEM = science, technology, engineering, and mathematics.

MEASURING CRITICAL THINKING

Critical thinking (CT) or problem solving has gained popularity as a concept in both educational and work settings. Many instruments in organizational and educational settings are said to measure critical thinking, and it has been proposed as an important skill for the workforce and, therefore, is an important training need in preparing future workers (Koenig, 2011). Here it is argued that as typically conceptualized, the research database supports a domain-specific skill conceptualization of critical thinking rather than the domain-general concept often discussed. A domain-general conceptualization argues that critical thinking is a skill that is applicable across settings and topics (thinking critically in general). In contrast, a domain-specific conceptualization posits that critical thinking is learned and applied to specific topics and content domains (thinking critically about educational policy vs. mechanical engineering). This distinction has large implications for the education and selection of the workforce. If critical thinking is a domain-general ability, then broad assessments of this ability will be useful for measuring the effectiveness of educational programs and personnel selection purposes. If what is called critical thinking is a catch-all for a large collection of specific knowledge and skill, some of which are job or field specific, then it cannot be assumed that education will lead to improvement across the board. Similarly, those skills that are of particular importance will need to be identified, measured, and trained. As a result, educational and personnel selection choices will need to be made.

A range of definitions have been proposed (Bangert-Drowns & Bankert, 1990; Ennis, 1989; Facione, 1990a, 1990b; Halpern, 1998) that range from very broad to more focused.

> Cognitive skills or strategies that increase the probability of a desirable outcome—in the long run, critical thinkers will have more desirable outcomes than "noncritical" thinkers. . . . Critical thinking is purposeful, reasoned, and goal-directed. It is the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions. (Halpern, 1998, pp. 450–451)

> Critical thinking, the ability and willingness to test the validity of propositions. (Bangert-Drowns & Bankert, 1990, p. 3)

> Critical thinking is reflective and reasonable thinking that is focused on deciding what to believe or do. (Ennis, 1985, p. 45)

> We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. (Facione, 1990b, p. 1)

Although these definitions differ in a number of ways, they generally focus on evaluating information and making decisions. Some are very broad (e.g., Halpern, 1998) to the point of arguably including all of problem solving, judgment, and cognition. Others are more specific (Bangert-Drowns & Bankert, 1990) and focus on a particular class of tasks.

Often a distinction is made in the CT literature between specific skills and what might be called *dispositions* or *attitudes*. Ultimately, this distinction is one of "can do" versus "will do." The former is the ability to correctly execute a CT skill, however defined. The latter is the interest and willingness to execute the skill. This review focuses primarily on the skill side.

The concepts of intelligence and expertise predate the study of critical thinking, and a critical scientific question is the extent to which CT is actually distinguishable from them. Definitions of both intelligence and expert performance overlap considerably with those for critical thinking. Simply on the basis of definitions, the overlap appears to be considerable. The definitions suggest that even if the concepts are independent, one may contribute to or be the developmental outcome of another. For example, intelligence might facilitate the development of expertise. Alternatively, CT might require some degree of intelligence.

Several operationalizations of CT exist that permit examination of key questions regarding the nature of CT. Measures of critical thinking include the Watson–Glaser Critical Thinking Appraisal, Cornell Critical Thinking Test, California Critical Thinking Skills Test, and the California Critical Thinking Disposition Inventory. A fairly sizable number of studies have examined relationships between each of these measures and both individual correlates and outcomes, with the majority of studies using the Watson–Glaser Critical Thinking Appraisal, California Critical Thinking Skills Test, and California Critical Thinking Disposition Inventory. Kuncel (2011) reviewed and synthesized this literature, which is reviewed next.

In Kuncel (2011), critical thinking measures generally exhibited moderate correlations with each other (the observed average correlation was .41) and slightly larger average correlations with traditional measures of cognitive ability (.48). Superstitious thinking and beliefs (often an important element in critical thinking definitions) were modestly correlated with both critical thinking scales (−.19) and traditional cognitive ability measures (−.13). This weak relationship is revisited in the Reasoning section discussing the newer construct, rationality. Finally, the relationship between critical thinking skills and the personality trait openness to experience was examined and had an average correlation of .24, very similar to the correlation observed between openness and general cognitive ability.

External correlates were also subjected to a meta-analysis for nursing performance. Critical thinking skill measures correlated with earned grades in school (.27), supervisory rating of job performance (.32), and nursing clinical decision making (.22). Overall, the results were similar but slightly weaker than what is observed for traditional cognitive ability measures. Moreover, the correlation between critical thinking skill measures and cognitive ability suggests that they would add little if any incremental validity.

Overall, the results of the meta-analysis suggested that existing critical thinking measures are not substantially different than general cognitive ability measures. However, before fully condemning the critical thinking measurement literature, a few caveats need to be mentioned. First, it is possible that critical thinking has been poorly measured by some scholars. However, a survey of the primary studies revealed little evidence that one measure is particularly more effective than others. Strong predictive evidence would need to be presented.

One area of promise is outside of the skill domain, identifying the disposition of skepticism and desire to consider, critique, and challenge statements. Although this disposition may be a combination of existing personality traits (it displays correlations with openness, e.g.), it may have important properties.

Second, the superiority of general cognitive ability measures over critical thinking measures is due, at least in part, to higher reliabilities. Many of the critical thinking measures are less reliable than, say, the SAT. If lengthened to increase their reliability, their predictive power would likely increase, although not enough to surpass general cognitive ability.

Finally, there is reason to believe that critical thinking is arguably an overarching term applied to a class of specific pieces of knowledge. Ennis (1985) proposed a list of specific critical thinking skills. This approach is the one that is likely to be the most productive for educating, selecting, and training the workforce.

This position is based on the literature finding that specific critical thinking type skills can be trained and can be made to generalize to a degree across settings but that critical thinking skills for one job or field are not necessarily those for other. Consider first the classic study presented in Nisbett, Fong, Lehman, and Cheng (1987), in which they presented research examining what they labeled statistical and methodological reasoning in graduate students in law, medicine, psychology, and chemistry. In both cross-sectional and longitudinal research, psychology students have been found to gain the most in these skills during graduate school, whereas chemistry graduate students have had little to zero gain.

Is it really possible that doctoral-level chemists do not learn how to reason? After all, on average, students applying for doctoral work in chemistry score appreciably higher than psychology doctoral

students on tests of quantitative reasoning. Careful examination of the Nisbett et al. (1987) study reveals a possible answer. The study tested students on statistical and methodological reasoning, namely, "methodological reasoning dealing with different types of confounded variable problems, for example, self-selection problems (26), sample bias problems" (Nisbett et al., 1987, p. 630). What is a self-selection effect in chemistry? These critical thinking skills are not of importance to the profession of chemistry (except perhaps analytical chemistry), and one might argue that students' time is better spent becoming expert chemists, not worrying about the problems of psychological research. Rather, the laws of thermodynamics are necessary for thinking at all effectively about an enormous range of problems from applied work in industry to theoretical problems. In contrast, psychology students do not learn these critical thinking skills, and one could reasonably conclude that psychology graduate students do not gain in their skill applying these natural laws to a range of problems during graduate training.

Note that this does not mean that the skills examined by Nisbett et al. (1987) are unimportant. They are very important for thinking critically about research or situations that involve self-selection effects or sampling bias. It might also be desirable for citizens and consumers of the news media for chemists to be more skilled at these questions. This point, though, relates to trade-offs between very specific skills rather than a discussion of the improving critical thinking for chemists.

REASONING

A growing body of research has examined what has been termed *rationality,* and it has been argued that reasoning is wholly or nearly independent from general cognitive ability. Reasoning research is largely anchored in a mixture of cognitive bias tasks (e.g., base rates) or other cognitive functioning tasks (argument evaluation, Wason Card Task, Iowa Gambling Task) used primary for research on people with psychopathology, brain lesions, or other impairments. Some of the research has yielded positive correlations with cognitive ability. For example, utility-maximizing response in probabilistic choice task was positively associated with cognitive ability (West & Stanovich, 2003). Similarly, those with higher cognitive ability made better use of correct Bayesian reasoning and performed better on deductive and inductive reasoning tasks (Stanovich & West, 1998). They also performed better, as a group, on the selection task, statistical reasoning assessments, and argument quality assessments (Stanovich & West, 1998). However, other studies have argued that some rationality tasks are effectively independent from general mental ability, and a theory has been proposed to account for differences in cognitive ability correlations with various reasoning tasks (Stanovich & West, 2008). These conclusions are interesting but may be premature for the following reasons.

Most of the tasks examined are single items at a single point in time collected in a laboratory setting in which student motivation is likely to be less than optimal, and using subjects who are restricted in range of cognitive ability. Moreover, many studies have used self-reported SAT scores, which are less reliable measures of cognitive ability than their actual scores (Kuncel, Crede, & Thomas, 2005). Simply put, the reliability of the measures and the ability to both detect effect and obtain accurate effect sizes are questionable. Studies that have treated a set of heuristic and biases tasks as a scale have revealed larger correlations with cognitive abilities (e.g., West & Stanovich, 2003). Despite these stronger correlations, which are consistent with increased reliability, this composite still demonstrated low reliability as indexed by internal consistency ($\alpha = .53$). Other studies using similar but psychometrically improved tasks have yielded observed correlations between a reasoning task and a *g* of .45 (Kaufman, DeYoung, Reis, & Gray, 2011). The magnitude of this effect is consistent with correlations observed for other more specific ability or knowledge measures.

Uncertain reliability also directly interferes with drawing clear conclusions. For example, research has observed modest correlations between a decision-making rationality task (Iowa Gambling Task) and traditional measures of cognitive ability (Toplak, Sorge, Benoit, West, & Stanovich, 2010). Toplak et al. (2010) concluded that the average

observed correlation is small enough to "highlight the separability between decision-making on the [Iowa Gambling Task] and cognitive abilities" (p. 562). However, no mention is made in the study of the Iowa Gambling Task's reliability. Without this information, it is not possible to draw firm conclusions about construct overlap. The observed effect could result in the opposite conclusion if the Iowa Gambling Task is of low reliability.

Missing from this literature on reasoning is a demonstration that the tasks under study are of any importance for the functioning of people in their regular lives. Such evidence is abundant for general cognitive ability but is not discussed for rationality tasks. If independence is clearly demonstrated for some of the tasks, it will be critical to prove that they are important as well. Independent and irrelevant is of limited psychological interest. The concern about importance is intimately linked to an ongoing concern about relevance of cognitive biases (e.g., Funder, 1987). Although errors may occur under some circumstances, the errors made by people may reflect compromises that are, on average, adaptive rather than maladaptive.

WORKING MEMORY CAPACITY

Working memory is a focal component of many information processing models of the mind. Perhaps the most well-known and influential model of working memory is Baddeley's model (Baddeley, 1986, 2000; Baddeley & Hitch, 1974). The general proposition is that instead of viewing memory as simple storage, it can be seen as a functional system that manipulates and maintains information. Baddeley and Hitch (1974) originally proposed a process with three components: the visuospatial sketchpad, the phonological loop, and the central executive. The visuospatial sketchpad and the phonological loop were theorized to deal with verbal–acoustic and visual–spatial information, respectively, and assist in the temporary storage and rehearsal of this information. The central executive process has a limited capacity and regulates the processing and integration of information from the other components. Although Baddeley (2000) has refined his model (adding an episodic buffer) over the years, and others have added to it (e.g., Oberauer, Süss, Schulze, Wilhelm, & Wittmann, 2000; Oberauer, Süss, Wilhelm, & Wittmann, 2003), the core concept is that of an active processor of incoming information.

Working memory capacity (WMC) is an individual difference variable derived from the broader theory of working memory. Modern theorists appear to view WMC in terms of working memory's central executive function and posit that it reflects attentional control in the face of distraction (e.g., Engle, 2002; Kane, Bleckley, Conway, & Engle, 2001), although Oberauer (2005) has suggested that this may be too broad. Over the past 2 decades, a debate has occurred on the fundamental nature of WMC, with some researchers claiming that WMC was essentially isomorphic (or the explanatory factor) to *g*, or fluid intelligence, because of the routinely high correlations found between the two in latent variable modeling (e.g., Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004; Kyllonen & Christal, 1990). Although some WMC researchers became skeptical of this conclusion (e.g., Conway, Kane, & Engle, 2003), the issue was brought to a head when Ackerman et al. (2005) presented the results of a meta-analysis that suggested that WMC and general mental ability were correlated much more modestly (average $\rho = .479$) than unity. Although responses by Oberauer, Schulze, Wilhelm, and Süss (2005) and Kane, Hambrick, and Conway (2005) disagreed with the size of the correlations and the tasks Ackerman et al. used to determine WMC, there seems to be general agreement that the two constructs are not identical.

Complex span tasks are the most common operationalization of WMC in behavioral psychology (Conway et al., 2005). In comparison to more simple span tasks that are designed to solely tap short-term recall, complex span tasks combine the presentation of stimuli with another cognitively demanding task that serves as a distraction. For instance, a reading span task might ask test takers to read a series of sentences out loud while remembering the last word from each of the sentences, then ask them to repeat these words in order (e.g., Daneman & Carpenter, 1980). There are many variations on this theme (e.g., verifying the veracity of the

sentences, using mathematical operations instead of sentence reading), but the tasks appear to be fairly reliable measures with respect to both internal consistency (Engle, Tuholski, Laughlin, & Conway, 1999; Kane et al., 2004) and test–retest reliability over at least a number of weeks (Conway et al., 2005). For a comprehensive introduction to complex span tasks and other WMC tasks, see Conway et al. (2005).

Although simple short-term memory tasks (e.g., digit span, digit symbol substitution) have been applied to work settings (Verive & McDaniel, 1996), it appears much less common with WMC tasks. As a result, little is known about the relationship between WMC and traditional work criteria. That being said, WMC seems potentially fruitful for the field of personnel selection.

First, it seems plausible that WMC could predict job performance with validity similar to that of typically used reasoning tests with the potential for less adverse impact. Building on theory from Jensen (1971), Verive and McDaniel (1996) presented the results of a meta-analysis for short-term memory tasks, showing that these simple tests had relatively high corrected validities for job performance and training (.41 and .49, respectively), with less than half the typical adverse impact found for general mental ability ($d = 0.42$). Given that most operationalizations of WMC combine short-term memory components and higher level executive functions, there is potential for measures of WMC to be in the middle ground of these two results. The potential for adverse impact reduction would likely be determined by how *g* loaded the specific WMC test used was. Regardless, because there is general agreement that the two constructs are not isomorphic, there should be some difference in relationships with work criteria when using WMC.

Second, preliminary data have suggested that WMC may be particularly important for tasks requiring multitasking, adaptive performance, or both (e.g., Pulakos, Arad, Donovan, & Plamondon, 2000). For instance, multiple studies (Bühner, König, Pick, & Krumm, 2006; Hambrick, Oswald, Darowski, Rench, & Brou, 2010; Hambrick et al., 2011; König, Bühner, & Mürling, 2005) have all reported that WMC was more related to performance on computer-based multitasking performance than was fluid intelligence and reasoning. Additionally, Oberlander, Oswald, Hambrick, and Jones (2007) found that WMC as measured by complex span tasks was negatively related to errors of commission (i.e., performing an incorrect action) when performing a multitasking simulation. Given that an element of multitasking is inherent in most operationalizations of WMC, these results are perhaps to be expected. To the extent that it is easier or more cost-effective to administer a test of WMC rather than the multitasking simulations themselves, they could be particularly useful in personnel selection settings in which a job analysis has indicated a high degree of multitasking along with other relevant constructs, such as polychronicity (e.g., Bluedorn, Kalliath, Strube, & Martin, 1999; König & Waller, 2010).

NEUROPSYCHOLOGICAL ASSESSMENTS

Previous research on neuropsychological assessments in work settings has most often focused on their association with return to work after brain trauma or in cases of other impairments (cognitive, psychopathology). However, a large class of carefully designed measures remain that assess a wide range of abilities. Many of these measures have been designed to directly assess very specific brain functions.

Although research on the predictive power of these measures is very limited, the results are intriguing and warrant additional attention. In a series of studies by Higgins, Peterson, Pihl, and Lee (2007), they examined the predictive power of dorsolateral prefrontal cognitive ability in combination with a more traditional measure of GMA and personality measures for both academic and work performance across several samples. In a work sample with a job of moderate complexity, the simple observed correlations for dorsolateral prefrontal cognitive ability were .42 to .57 with supervisory ratings. In a second sample for a lower complexity job, the relationships were more far more modest ($r = -.12$). Dorsolateral prefrontal cognitive ability yielded incremental predictive power above and beyond GMA and personality in the academic samples in which simple correlations were .37 and .33

(a GMA measure was not included in the work samples). Similar results in an academic setting were reported in Peterson, Pihl, Higgins, Seguin, and Tremblay (2003). These studies, although based on relatively small sample sizes around 100, have suggested a direction of potential interest to industrial and organizational psychology. The academic performance results are most compelling for prediction of training outcomes, whereas the work results, combined with the multitasking research reviewed in this chapter, have suggested promise for high-complexity jobs.

CONCLUSION

Decades of research on human cognitive abilities has established several consistent findings. Cognitive abilities appear to fit into a hierarchic structure with an overarching general factor. Cognitive abilities are also consistently good predictors of academic performance, job training, and job performance, yet prediction of these same outcomes can be further improved by measuring aspects of temperament and interest. These results provide a strong foundation for the field. Future research in work settings should consider developmental, neuropsychological, and information processing approaches. Even if these approaches do not immediately yield improved prediction, they hold promise for improving industrial and organizational psychologists' understanding of learning and work and may provide new foundations for the future.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3–27. doi:10.1037/0033-2909.102.1.3

Arneson, J. J., Sackett, P. R., & Beatty, A. S. (2011). Ability performance relationships in education and employment settings: Critical tests of the more-is-better and the good-enough hypothesis. *Psychological Science, 22*, 1336–1342. doi:10.1177/0956797611417004

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. D. (1986). *Working memory*. Oxford, England: Clarendon Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Bangert-Drowns, R. L., & Bankert, E. (1990, April). *Meta-analysis of effects of explicit instruction for critical thinking*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Bemis, S. E. (1968). Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology, 52*, 240–244. doi:10.1037/h0025733

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science, 20*, 822–830. doi:10.1111/j.1467-9280.2009.02368.x

Bluedorn, A. C., Kalliath, T. J., Strube, M. J., & Martin, G. D. (1999). Polychronicity and the Inventory of Polychronic Values (IPV): The development of an instrument to measure a fundamental dimension of organizational culture. *Journal of Managerial Psychology, 14*, 205–231. doi:10.1108/02683949910263747

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line*. Princeton, NJ: Princeton University Press.

Bühner, M., König, C. J., Pick, M., & Krumm, S. (2006). Working memory dimensions as differential predictors of the speed and error aspect of multitasking performance. *Human Performance, 19*, 253–275. doi:10.1207/s15327043hup1903_4

Campbell, J. P., & Kuncel, N. R. (2001). Individual and team training. In N. Anderson, D. S. Ones, H. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology: Vol. 1. Personnel psychology* (pp. 278–312). London, England: Sage. doi:10.4135/9781848608320.n14

Carroll, J. B. (1974). The aptitude–achievement distinction: The case of foreign language aptitude and proficiency. In D. R. Green (Ed.), *The aptitude–achievement distinction* (pp. 286–303). Monterey, CA: CTB/McGraw-Hill.

Carroll, J. B. (1993). *Human cognitive abilities*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511571312

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38*, 592.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston, MA: Houghton Mifflin.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30*, 15–41. doi:10.1037/0003-066X.30.1.15

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence, 32*, 277–296. doi:10.1016/j.intell.2003.12.002

Colquitt, J. A., LePine, J. A., & Noe, R. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*, 678–707. doi:10.1037/0021-9010.85.5.678

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review, 12*, 769–786. doi:10.3758/BF03196772

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7*, 547–552. doi:10.1016/j.tics.2003.10.005

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability–performance relationships: A reconfirmation. *Journal of Applied Psychology, 75*, 297–300. doi:10.1037/0021-9010.75.3.297

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and learning. *Journal of Verbal Learning and Verbal Behavior, 19*, 450–466. doi:10.1016/S0022-5371(80)90312-6

Defense Manpower Data Center. (2006). *CAT–ASVAB Forms 1 and 2* (Technical Bulletin No. 1). Seaside, CA: Author.

Drasgow, F. (2002). Intelligence and the workplace. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 107–130). New York, NY: Wiley.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19–23. doi:10.1111/1467-8721.00160

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128*, 309–331. doi:10.1037/0096-3445.128.3.309

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*, 44–48.

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*, 4–10. doi:10.2307/1174885

Facione, P. A. (1990a). *California Critical Thinking Skills Test manual*. Millbrae: California Academic Press.

Facione, P. A. (1990b). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction, research findings, and recommendations*. Newark, DE: American Philosophical Association.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*, 75–90. doi:10.1037/0033-2909.101.1.75

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449–455. doi:10.1037/0003-066X.53.4.449

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology, 24*, 1149–1167. doi:10.1002/acp.1624

Hambrick, D. Z., Rench, E. M., Poposki, E. S., Darowski, D. R., Bearden, R. M., Oswald, F. L., & Brou, R. (2011). The relationship between the ASVAB and multitasking in Navy sailors: A process-specific approach. *Military Psychology, 23*, 365–380.

Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology, 93*, 298–319. doi:10.1037/0022-3514.93.2.298

Humphreys, L. G. (1984). General intelligence. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing*. New York, NY: Plenum Press.

Hunt, E. (2011). *Human intelligence*. New York, NY: Cambridge University Press.

Jensen, A. R. (1971). Do schools cheat minority children? *Educational Research, 14*, 3–28. doi:10.1080/00131881710140101

Johnson, W., & Bouchard, T. J., Jr. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 393–416. doi:10.1016/j.intell.2004.12.002

Johnson, W., Bouchard, T. J., Jr., McGue, M., Segal, N. L., Tellegen, A., Keyes, M., & Gottesman, I. I. (2007). Genetic and environmental influences on the verbal-perceptual-image rotation (VPR) model of the structure of mental abilities in the Minnesota study of twins reared apart. *Intelligence, 35*, 542–562. doi:10.1016/j.intell.2006.10.003

Judge, T. A., Colbert, A. E., & Ilies, R. (2004). Intelligence and leadership: A quantitative review and test of theoretical propositions. *Journal of Applied Psychology, 89*, 542–552. doi:10.1037/0021-9010.89.3.542

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*, 169–183. doi:10.1037/0096-3445.130.2.169

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 66–71. doi:10.1037/0033-2909.131.1.66

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*, 189–217. doi:10.1037/0096-3445.133.2.189

Kaufman, S. B., DeYoung, C. G., Reis, D. L., & Gray, J. R. (2011). General intelligence predicts reasoning ability even for evolutionarily familiar content. *Intelligence, 39*, 311–322. doi:10.1016/j.intell.2011.05.002

Koenig, J. A. (Ed.). (2011). *Assessing 21st century skills*. Washington, DC: National Academies Press.

König, C. J., Bühner, M., & Mürling, G. (2005). Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and extraversion are not. *Human Performance, 18*, 243–266. doi:10.1207/s15327043hup1803_3

König, C. J., & Waller, M. J. (2010). Time for reflection: A critical examination of polychronicity. *Human Performance, 23*, 173–190. doi:10.1080/08959281003621703

Kuncel, N. R. (2011). Measurement and meaning of critical thinking. In J. A. Koenig (Ed.), *Assessing 21st-century skills*. Washington, DC: National Academies Press.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis. *Review of Educational Research, 75*, 63–82. doi:10.3102/00346543075001063

Kuncel, N. R., & Hezlett, S. A. (2007a). Standardized tests predict graduate student success. *Science, 315*, 1080–1081. doi:10.1126/science.1136618

Kuncel, N. R., & Hezlett, S. A. (2007b). Utility of standardized tests: Response. *Science, 316*, 1696–1697.

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19*, 339–345. doi:10.1177/0963721410389459

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162–181. doi:10.1037/0033-2909.127.1.162

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86,* 148–161.

Kuncel, N. R., & Klieger, D. M. (2012). Predictive bias in work and educational settings. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 462–484). Oxford, England: Oxford University Press.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*, 389–433. doi:10.1016/S0160-2896(05)80012-1

Lubinski, D. (2009). Exceptional cognitive ability: The phenotype. *Behavior Genetics, 39*, 350–358. doi:10.1007/s10519-009-9273-0

Lubinski, D., & Dawes, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence*. Baltimore, MD: Williams & Wilkins.

Mattern, K. D., & Patterson, B. F. (2009). *Is performance on the SAT related to college retention?* (College Board Research Report 2009–07). New York, NY: College Board.

McCloy, R. A., Russell, T. L., & Wise, L. L. (1996). *GATB improvement project final report*. Washington, DC: U.S. Department of Labor.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335–354. doi:10.1111/j.1744-6570.1990.tb01562.x

Muchinsky, P. M. (1993). Validation of intelligence and mechanical aptitude tests in selecting employees for manufacturing jobs. *Journal of Business and Psychology, 7,* 373–382. doi:10.1007/BF01013752

Murphy, K. R., Dzieweczynski, J. L., & Yang, Z. (2009). Positive manifold limits the relevance of content matching strategies for validating selection test batteries. *Journal of Applied Psychology, 94*, 1018–1031. doi:10.1037/a0014075

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101. doi:10.1037/0003-066X.51.2.77

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238*, 625–631. doi:10.1126/science.3672116

Oberauer, K. (2005). The measurement of working memory capacity. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 393–408). Thousand Oaks, CA: Sage. doi:10.4135/9781452233529.n22

Oberauer, K., Schulze, R., Wilhelm, O., & Süss, H.-M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*, 61–65. doi:10.1037/0033-2909.131.1.61

Oberauer, K., Süss, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—Facets of a cognitive ability construct. *Personality and Individual Differences, 29*, 1017–1045. doi:10.1016/S0191-8869(99)00251-2

Oberauer, K., Süss, H.-M., Wilhelm, O., & Wittmann, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31*, 167–193. doi:10.1016/S0160-2896(02)00115-0

Oberlander, E. M., Oswald, F. L., Hambrick, D. Z., & Jones, L. A. (2007). *Individual difference variables as predictors of error during multitasking* (Report No. NPRST-TN-07–9). Millington, TN: Navy Personnel Research, Studies, and Technology.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in personnel selection decisions. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 143–173). Oxford, England: Blackwell.

Peterson, J. B., Pihl, R. O., Higgins, D. M., Seguin, J. R., & Tremblay, R. E. (2003). Neuropsychological performance, IQ, personality, and grades in a longitudinal grade-school male sample. *Individual Differences Research, 1*, 159–172.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the work place: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*, 612–624. doi:10.1037/0021-9010.85.4.612

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and postsecondary academic performance? *Psychological Bulletin, 135*, 1–22.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88*, 1068–1081. doi:10.1037/0021-9010.88.6.1068

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*, 187–210.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274. doi:10.1037/0033-2909.124.2.262

Shen, W., Sackett, P. R., Kuncel, N. R., Beatty, A. S., Rigdon, J. L., & Kiger, T. B. (2012). All validities are not created equal: Determinant in variation in SAT validity across schools. *Applied Measurement in Education, 25*, 197–219.

Silva, J. M., & White, L. A. (1993). Relation of cognitive aptitudes to success in foreign language training. *Military Psychology, 5*, 79–93. doi:10.1207/s15327876mp0502_1

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161–188. doi:10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672–695. doi:10.1037/0022-3514.94.4.672

Stanton, H. M., Koerth, W., & Seashore, C. E. (1930). *Musical capacity measures of adults repeated after music education: A study of retest scores and ratings in the Seashore measures of musical talent after three years of music education in the Eastman School of Music* (University of Iowa Studies, First Series No. 189). Iowa City: University of Iowa.

Toplak, M. E., Sorge, G. B., Benoit, A., West, R. F., & Stanovich, K. E. (2010). Decision making and cognitive abilities: A review of associations between Iowa Gambling Task performance, executive functions, and intelligence. *Clinical Psychology Review, 30*, 562–581. doi:10.1016/j.cpr.2010.04.002

Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23*, 15–32. doi:10.1016/S0160-2896(96)80003-1

West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associates between probabilistic choices and cognitive ability. *Memory and Cognition, 31*, 243–251. doi:10.3758/BF03194383

Zeidner, J., & Johnson, C. D. (1994). Is personnel classification a concept whose time has passed? In M. G. Rumsey, C. B. Valker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 377–410). Hillsdale, NJ: Erlbaum.

CHAPTER 25

BIOGRAPHICAL INFORMATION

Neal Schmitt and Juliya Golubovich

In industrial and organizational psychology, it is taken as a truism that one's past behavior is the best predictor of one's future behavior. This notion has a long history in the research on individual differences (Allport, 1937; Galton, 1902). What is treated as biographical data or, alternatively, "biodata," has changed since it was first introduced as a method of measurement. Originally, it represented relatively objective items that addressed one's background (e.g., education level, gender, ethnicity, address, previous jobs, number of siblings). Responses to these items were differentially weighted and scored relative to their relationship with some outside criterion an investigator was interested in predicting (England, 1971). Biodata have evolved to include items addressing one's hobbies, interests, recreational preferences, educational and job preferences, experiences, and self-appraisals (Hough & Paullin, 1994). Examples of four items used to measure social responsibility among college students are reproduced in Exhibit 25.1. The first two of these items are relatively objective, whereas the second and third are similar to the latter type of item just described and are similar to items that might appear in a personality test, as is often the case in more recently used measures of biodata. When biodata are discussed in this chapter, the broader, more inclusive definition of that term is used unless specifically stated otherwise. Although this approach to measurement has been used most frequently in work and educational contexts, examples of its use in other contexts is provided as well (e.g., social, clinical, and developmental psychology).

As evidence of this notion that biographical differences capture enduring differences between people, Ghiselli (1966) and Henry (1966) found that biodata were one of the best predictors of a variety of work performance outcomes. England (1971) provided what is probably the best description of the manner in which early researchers used biodata (or *weighted application blanks,* as England called them). In early applications, researchers took the items on a typical application blank, occasionally adding items of particular interest, and developed a set of weights for each response that usually reflected the items' presumed relationship to some organizationally relevant outcome (e.g., turnover). The details of these weighting schemes are described later, but their most important characteristic was that the weights were empirically determined at the level of a response alternative.

Although this empirical approach produced a high level of criterion-related validity, such validity was often specific to the occupation or organization in which the instrument was developed. Biodata items chosen and scored empirically tended to lack generalizability across time and situations. Mitchell and Klimoski (1982) provided contradictory evidence, however, in that empirically derived scales provided slightly better cross-validated validities than rationally derived scales. Also, evidence has shown that empirically derived scales may prove valid across various situations if developed to tap "attributes of interpersonal behavior that . . . possess an integral relationship to all forms of social interaction" (Gough, 1968, as cited by McIntyre, Mauger,

DOI: 10.1037/14047-025
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

Exhibit 25.1
Four Items Written to Represent Social Responsibility in College Students

In the past year, how many hours of volunteer work did you perform?
a. 0
b. Between 1 and 10
c. Between 11 and 30
d. Between 31 and 75
e. More than 75

During the past year, how many times have you given money, food, or clothes to a charity or poor person in need?
a. 0
b. 1
c. 2
d. 3
e. More than 3 times

How likely are you to pick up litter that you come across and carry it until you find a trash receptacle?
a. Extremely likely
b. Very likely
c. Somewhat likely
d. Not very likely
e. Unlikely or not at all

How important has it been in the past year for you to be involved in community or volunteer work?
a. Extremely important
b. Very important
c. Important
d. Not very important
e. Not at all important

Note. All of these items were scored on a continuum; that is, more hours or more times are considered to be better.

Margalit, & Figueiredo, 1989, p. 385). This was the case with the Socialization Scale of the California Psychological Inventory, which, in addition to some traditional biodata items, includes a large number of more personality-like items (it is, in fact, considered a personality scale). Not surprisingly, it has demonstrated (a) good generalizability to various groups (it was originally developed using delinquent adolescents; Kadden, Litt, Donovan, & Cooney, 1996) and (b) cross-cultural validity (Schalling, 1978, as cited by Standage, Smith, & Norman, 1988). Similarly, in other instances in which scored biodata were found to generalize, the items were often written to minimize situational peculiarities (e.g., Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990). However, even in the presence of comparable or superior validity, it has often been difficult to ascertain the constructs measured by these instruments and provide a theoretical or intuitive reason for the validity of empirically derived scales.

The highly empirical approach represented by the work of England (1971) continued until Owens (1976; Owens & Schoenfeldt, 1979) and Mumford (Mumford, Costanza, Connelly, & Johnson, 1996; Mumford & Stokes, 1992) espoused and demonstrated a more rational (theoretical) and much expanded approach to the use of biodata. Owens and Schoenfeldt (1979) felt that the differences among people result partly from developmental patterns arising from major life-history experiences. They clustered undergraduate students on the basis of their profile of responses to a large battery of biodata items. The Owens and Schoenfeldt biodata items captured important behaviors and experiences related to student development. Using scores derived from principal-components analyses of responses to the biodata items, they clustered 2,000 freshmen into 23 male and 15 female subgroups. To evaluate this subgroup approach, Owens and Schoenfeldt assessed the degree to which subgroup membership was associated with external performance criteria. Subgroup status was related to a variety of educational outcomes including over- and underachievement, college GPA, academic probations and dismissals, and a number of course withdrawals in a series of master's theses and dissertations. Mumford, Connelly, and Clifton (1990) reported that in addition to performance, subgroups identified in this manner also predicted motivational criteria such as person–job fit and situational choice.

This largely post hoc attempt to derive meaning from a large body of biodata responses has been followed by multiple recent attempts to build specific constructs into the measures during the item-writing or item-generation process (Mumford, 1999). For example, Karas and West (1999) developed biodata items for applicants to Australian Public Service jobs to measure six constructs: goal orientation, teamwork, customer service, resourcefulness,

learning ability, and leadership. These constructs explained the applicants' responses, but empirically derived scoring keys did better than rational scoring in terms of predicting job performance ratings. Schmitt, Jennings, and Toney (1999) provided confirmation of the existence of three of six a priori constructs in a biodata battery designed to measure applicants' potential for a law enforcement position. Schoenfeldt (1999) reported more positive results with respect to utility of rational scales. Three of five performance criteria on rationally derived scales for salesperson jobs showed validities in the .30s, whereas empirical scales failed to cross-validate. In yet another example, Schmitt et al. (2009) reported good internal consistency reliability and discriminant validity for a set of 11 rationally derived biodata scales used to measure college student success. Mumford et al. (1996) described a series of studies directed at the measurement of a variety of constructs relevant to several different jobs or organizations and provided evidence pertaining to the internal consistency and validity of the measures.

The history of the use of biodata has progressed through four stages: First, a simple recognition that information in an application blank could be used to predict subsequent job behavior; second, a systematic weighting of application blank items and responses by virtue of their relationships with some external criterion; third, a stage in which post hoc empirical identification of clusters of people with similar biographical profiles was used to interpret the developmental meaning of these profiles; and finally, a move to a priori theoretical consideration of the constructs in which there is interest and the generation and evaluation of biodata items that should reflect those constructs. In the following pages, the manner in which items are generated, the variety of item formats and scoring that have been used, and the data regarding validity and utility are described. The conclusion provides productive future avenues of research and practical advice.

ITEM GENERATION

Biodata are most typically used in selecting individuals into jobs (Mumford, 1999), and thus, job performance will typically serve as a starting point for determining what information biodata items need to elicit (Mumford et al., 1996). With a job-oriented strategy, items are written to get at past behaviors that are either direct antecedents or manifestations of the criterion of interest (e.g., job requirements biodata; Allworth & Hesketh, 2000; Stokes, Toth, Searcy, Stroupe, & Carter, 1999). With a worker-oriented approach, items are written to get at the constructs believed to underlie the criterion of interest (construct-oriented biodata; Allworth & Hesketh, 2000; Colangelo, Kerr, Hallowell, Huesman, & Gaeth, 1992; Kilcullen, White, Mumford, & Mack, 1995). A worker-oriented approach is more appropriate than a job-oriented approach when the applicant pool is expected to have had limited opportunity to deal with the types of situations to be encountered on the job (Mumford et al., 1996). Without direct experience with the job, responses to job-oriented items would have to be negative for many or most items; if one uses a worker-oriented approach, items reflecting experience or interest in activities (presumably reflecting particular skills, abilities, or motivation) in other domains of life can be used. If one examines the four items in Exhibit 25.1, none of them reflect activities related to social responsibility in the academic arena; the hypothesis is that engaging in such activities in other spheres of life will be reflected in social responsibility in the college environment as well (e.g., reflected perhaps in tutoring less able students, involving oneself in student government).

The goal of the worker-oriented approach is to assess differential expression of performance-relevant constructs (knowledge, skills, abilities, other characteristics) in various situations. Those taking this approach need to first determine the performance-relevant constructs. A job analysis, performance appraisal instrument (e.g., Russell, Matson, Devlin, & Atwater, 1990), or theory (e.g., Mumford et al., 1996) would provide this information. Second, situations that would have elicited the performance-relevant constructs need to be identified. To do so, the types of experiences that are relevant for the constructs of interest need to be considered, and determining which of these experiences the target sample is likely to have had is necessary. Mumford (1999) suggested that situations

should not be restricted to routine ones only; less typical, high-demand situations (i.e., critical incidents; Flanagan, 1954) could prove especially useful. The third step would be to write items that will capture differential expression of the performance-relevant constructs by determining how individuals interpreted and behaved in and what they gained from the situations in question (Mumford et al., 1996; Russell, 1994). When writing items, it is important to choose situations that are not so strong that they would have precluded the expression of individual differences in the construct of interest (Mumford et al., 1996). For instance, most people would react emotionally to the death of a friend; however, the range of emotional reaction to a friend's failure to receive a promotion in her or his workplace might be greater.

There are multiple sources to turn to when actually coming up with biodata items. Investigators can rely on their own psychological knowledge (Mumford & Owens, 1987), or they might assemble a panel of diverse, well-educated, and psychometrically trained individuals for the item-generation task (e.g., Mumford et al., 1996). Notably, Russell (1994) suggested that relying on investigators' knowledge to generate items may not be an adequate strategy in that it can result in deficiency or contamination of measures, and they proposed letting incumbents, who are in a sense subject matter experts, assist with the task by writing life history essays or responding to interview-type questions (see Colangelo et al., 1992, and Russell et al., 1990, for examples). After defining the critical performance dimensions, one would strategically (so as to elicit developmental experiences relevant to the critical performance dimensions) choose topics about which to query incumbents. The responses would be content analyzed to generate large sets of biodata items.

Another potential source of biodata items are theories of adult development (Mumford, 1999) or of individual differences (e.g., personality, vocational choice, leadership; Russell, 1994). Items based on theory-guided hypotheses about dimensions underlying performance are expected to have strong relationships with performance criteria (Nickels, 1994). However, it is important that the item pool to adequately cover the predictor space, not include irrelevant behaviors and experiences, and capture experiences that would have been accessible to different demographic groups (e.g., age, race, gender; Mumford & Owens, 1987). Starting off with an instrument previously used by others and adapting it (e.g., Collins & Schmidt, 1993; Mael, 1995) or selecting relevant items from existing biodata banks (e.g., Schmitt, Oswald, Kim, Gillespie, & Ramsay, 2003; Whitney & Schmitt, 1997) are alternatives to generating one's own items.

TYPES OF ITEMS

So as to better understand why and how biodata are able to predict various criteria, work has been done to differentiate items along various dimensions (Lefkowitz, Gebbia, Balsam, & Dunn, 1999). As Lefkowitz et al. (1999) summarized in their review of that work, differentiation has been based on observed content domains (e.g., school achievement, interests), item structure or format (e.g., number and style of response options), and implicit attributes (e.g., verifiability, transparency).

Content Domains

Biodata items typically cover some of the following domains: personal, general background, education, employment experience, skills, socioeconomic status, social, interests, and personal characteristics and attitudes (Crafts, 1991). The decision of which domains should be covered is typically based on the dimensions of the criterion that the instrument will be designed to predict. However, other considerations may come into play as well. For example, Mumford and Owens (1987) reviewed some evidence suggesting that different biodata domains may be differentially susceptible to socially desirable responding.

Item Format

With regard to item format, multiple-choice (i.e., respondents select one of the provided options; see Items 1 and 2 in Exhibit 25.1) and Likert-type formats (e.g., respondents indicate on a scale the extent to which their previous job or experience required use of a particular skill, as is the case for Items 3 and 4 in Exhibit 25.1) are most typically used. Researchers

have suggested formatting multiple-choice response options so that they have a neutral or positive connotation, form a continuum if applicable, and provide an escape option if all possible responses options are not included (Mumford & Owens, 1987). The advice to keep all items neutral or positive is tempered, however, by findings showing that negative items are more indicative than positive items of certain constructs of interest (e.g., neuroticism; Reiter-Palmon, DeFilippo, & Mumford, 1990). Also, the developmental role that learning to deal effectively with negative life events plays in life histories may be an important determinant of subsequent behavior (Dean, Russell, & Muchinsky, 1999).

Some have used a dichotomous response format (e.g., yes–no; Van Iddekinge, Eidson, Kudisch, & Goldblatt, 2003) for biodata items, but this format restricts item variance and may limit the criterion-related validity of the instrument. Forced-choice formats are another alternative to the more frequently used formats discussed earlier. Relative to Likert-type response formats, forced-choice formats are considered more resistant to faking but result in artificially lower correlations between constructs and also restrict validity, as has been demonstrated by Hicks (1970). Snell, Sydell, and Lueke (1999) also alluded to these measurement difficulties when the forced-choice format produces ipsative scales (i.e., high scores on one scale require that a respondent receive low scores on another scale). Another possibility is for biodata to be collected via an essay format in which individuals are prompted to describe how they acted in a particular situation, for example. This format seems especially well suited to evaluating motivation and problem-solving processes (Mumford, 1999). The validity of this format has also been evidenced when it comes to predicting performance (Hough, 1984). Advantages of the essay format pertain to applicant reactions and faking. This format may get a more favorable reception from applicants because it allows more freedom than a more structured format for self-description, and it may also be less susceptible to faking than a multiple-choice format (Mumford, 1999). The fact that with the essay format more work has to be done on the scoring end (e.g., ratings along a set of dimensions, content analysis) may be seen as a drawback. In addition, the writing requirement likely means that verbal ability is a correlate of biodata information collected in this manner, which may be a liability in some applications.

Implicit Attributes

Mael (1991) laid out a 10-dimension taxonomy of biodata's implicit attributes. He asserted that the only necessary attribute of biodata items is that they be historical (as opposed to future oriented or hypothetical); more is left to discretion with other item attributes. Decisions are typically guided by concerns about effects on applicants' ability to fake, potential legal repercussions of asking applicants inappropriate questions, and ethical considerations. Findings have indicated that lower objectivity (objective items require recall of information; no subsequent interpretation of information of that information is needed), lower verifiability (verifiable items are ones for which responses can be checked using sources other than the word of the respondent), lower discreteness (items that ask about a single behavior or count of behaviors as opposed to a summary measure of multiple instances of a behavior—e.g., average time spent—are considered discrete), and lower externality (external items ask about external actions as opposed to internal thoughts, reactions, and attitudes) are associated with more faking (Becker & Colquitt, 1992), as evidenced by higher scores on a biodata measure. If the four social responsibility items in Exhibit 25.1 are considered, the first two items would be considered fairly objective, verifiable, discrete, and external. Items 3 and 4, however, would qualify as subjective, nonverifiable, and internal. Also, the discreteness dimension would not apply in the case of these two items.

Higher transparency may also increase susceptibility to faking—being able to recognize what is being measured should make it easier to fake items in the appropriate direction (Snell et al., 1999). Related to this, faking may be more difficult when applicants cannot figure out how item responses will be scored (i.e., when it is hard to determine which response is the most desirable; Snell et al., 1999). However, researchers have not accumulated strong evidence to show that transparent items are more

fakeable than more subtle items (Hough & Paullin, 1994). Actually, it appears that given enough motivation, items of all types can be faked. The climate (e.g., level of test security, coaching practices) in which the instrument is used may also influence the prevalence of faking (Mael, 1991).

With regard to legal considerations, items that may be more resistant to faking are also ones that may be inadvisable to ask from a legal standpoint (e.g., they are less transparent and consequently less face valid; Mael, 1991). Finally, ethical considerations can also be brought to bear on the issue of item attributes. Some may see items that are noncontrollable (items concerning things that happened to or were done to an individual rather than actions one chooses to do or not do) and not equally accessible (items dealing with skills and experiences that not all applicants will have had access to) as unfair because individuals may score poorly on those items (if they respond honestly) for reasons that are not fully under their control (Mael, 1991). Clearly, trade-offs must be negotiated when designing items. The social responsibility items in Exhibit 25.1 are arguably controllable but perhaps not equally accessible in that individuals of lower socioeconomic status would have less opportunity to volunteer their resources (e.g., time, money, food), and it may not make sense for these individuals to try to clean up litter if litter is a rampant problem where they live.

METHODS OF SCORING BIODATA

Once a pool of items has been generated (and screened to ensure appropriateness of the items), one must decide how best to score items so as to allow for prediction of the criterion of interest on the basis of scores on the biodata measure. Researchers have used three major ways of developing scores based on biodata: empirical keying, rational scoring based on the content of the items, or a scoring system derived from a factor or cluster analysis of items.

Empirical Keying

As mentioned in the first section of this chapter, recognition of the potential of biographical information as an index of human potential was followed by an effort to weight responses to biodata items so as to maximize their relationship to some outcome of interest (England, 1971). Steps taken in the development of a scoring key using this approach are as follows:

1. Choose an appropriate outcome measure (e.g., turnover, accidents, and performance).
2. Identify the status of a group of individuals on this outcome and define outcome groups (e.g., low and high performers; those who turn over and those who stay).
3. Collect responses to biodata items.
4. Determine the responses of each of the outcome groups.
5. Identify response option differences between the outcome groups and develop the scoring key. That is, assign higher values to options chosen by individuals scoring on the desirable end of the criterion of interest (e.g., low on turnover, high on performance)
6. Check the relationship of the set of scored biodata items with the outcome measures on a cross-validation group.

An example of this approach to scoring is represented in Table 25.1. In this case, the second and third response alternatives are scored positively, reflecting the fact that individuals who endorsed these two options were more likely to complete a training course that was the criterion or outcome of interest. The first and fourth alternatives were scored negatively because individuals who chose these two options were more likely to leave the training course before completion. Larger differences in the two outcome groups were reflected in scores of 2 or −2 as opposed to smaller differences, which were assigned scores of 1 or −1. A zero score was assigned to the fifth option because the two outcome groups showed a very small difference in endorsement of this option. The magnitude of the difference justifying differential scoring is sometimes determined by a statistical significance test but more often determined by some arbitrary judgment about the magnitude of the difference that justifies differential scoring. Obviously, this method of determining scores is subject to sample variations, so a check (i.e., cross-validation) as to whether the

TABLE 25.1

Example of Responses to a Biodata Item ("When Reading for Pleasure, What Type of Literature Are You Most Likely to Read?") and the Corresponding Scoring Key

Response option	Proportion of group who leave during training (%)	Proportion of group who complete training (%)	Scoring key
Adventure stories	35	10	−2
Historical novels or biographies	20	45	2
Books about science	3	15	1
Mysteries or love stories	28	19	−1
Science fiction or horror	14	11	0

scoring key does indeed predict training completion in another sample from the same group of trainees would be essential.

Rational Scoring

Because of the atheoretical nature of empirical scoring methods, researchers began using a more rational approach to scoring. This approach began with identification of the constructs thought to be important determinants of a criterion either through job analyses or theory. Items and response options that are thought to reflect these constructs are then generated and scored consistent with the a priori ideas about the criterion. A great many attempts have been made to develop rationally derived and scored biodata instruments that follow some variation of these general steps (e.g., Burisch, 1984; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Schmitt et al., 1999; Stokes et al., 1999). Mumford (1999) has argued strongly that if researchers are to benefit theoretically and practically from the use of biodata, a substantive or rational approach to construction of biodata scales must be adopted. Oswald, Schmitt, Kim, Gillespie, and Ramsay (2004) provided a typical example of the rational development of biodata scales to measure college student performance. The steps they used were (a) a content analysis of the dimensions of student performance as identified in the university's mission or goal statements; (b) writing and selection of items that were judged to be indicators of performance on these dimensions; (c) scoring of item responses on the basis of what the authors judged to be good answers (these biodata item response scales are usually ordinal in nature, so scoring is usually continuous in nature); and (d) collection of response data, computation of scale reliabilities and intercorrelations, and confirmatory factor analyses to test the presumed structure of the biodata items and scales. An example of four items designed to measure social responsibility is presented in Exhibit 25.1. Multiple judges agreed that these four items were indicators of the standing of student participants on a social responsibility dimension.

Disagreement certainly remains among measurement experts as to the value of rational item development and scoring. As mentioned earlier in the chapter, some researchers have found the empirical approach superior to the rational approach (e.g., Mitchell & Klimoski, 1982). It is also the case that items originally developed using an empirical approach acquire scientific "respectability" as constructs over time with the accumulation of research data. Today's interpretations of Minnesota Multiphasic Personality Inventory scores are one good example. Our own view is that a rational approach will more rapidly lead to ease of interpretation and connectedness to the broader scientific knowledge base, although no data to support this assertion probably exist.

Factor-Analytic or Internal Approach

The third approach to developing and scoring biodata instruments is to use factor or cluster analysis to "discover" the dimensions underlying responses to a set of items. This approach is often also called an *internal approach* to scale development; it was probably first used by Owens and colleagues

(e.g., Mumford & Owens, 1984; Owens, 1976; Owens & Schoenfeldt, 1979) as a basis for clustering students into groups who had similar backgrounds or developmental histories. In developing scoring methods using this approach, the researcher writes or selects a large number of biodata items that are considered relevant to some domain of interest (e.g., citizenship behavior in an organization). Response scales are developed for each item that represents a practical continuum given the targeted respondent (e.g., 1–5 leadership positions as opposed to 1–500 such positions over a year's time frame). These items are then administered to a large group, and responses are typically factor or cluster analyzed to discover the underlying dimensions. Items deemed to belong to these dimensions are then assigned to the same scale and scored on that dimension. The content of the items in each scale is examined to provide interpretive context. Further psychometric data may be collected or generated, such as coefficient alpha for the items in each scale, scale intercorrelations to assess the degree to which groups of items or scales do measure different constructs, and correlations with external measures such as performance, turnover, or other behaviors of interest to assess construct validity. Aside from the limitations associated with factor analysis of these measures (see Hough & Paullin, 1994, pp. 112–113), one also has to be concerned about whether the original set of items is representative of the domain of interest. A factor analysis (or any other data-analytic approach for that matter) cannot reveal a factor or dimension if it is not represented in the original set of items. However, it is the case that this inductive factor-analytic approach can produce internally consistent and interpretable dimensions that facilitate theoretical development of how past experience determines future behavior (Mumford, 1992).

Comparisons of Scoring Approaches

Several studies, beginning with Mitchell and Klimoski (1982), have compared different approaches to the development of biodata scores. Mitchell and Klimoski found that cross-validated empirical biodata scoring keys produced superior predictions of the attainment of a real estate license by real estate students when compared with predictions from scales produced by an internal analysis of responses as described in the previous section of this chapter. Hough and Paullin (1994) provided a review of 21 studies that allowed for comparisons of two or more of the approaches to biodata scales described earlier. They found that the validity of the rational approach exceeded that of the empirical approach by .01 across 14 such comparisons, and the validity of the rational approach was higher on average than that of the internal or factor-analytic approach by .05 in 12 comparisons. In 16 comparisons of the empirical and internal approach, the validity of the empirical approach exceeded that of the internal approach by .02. So, although there do not appear to be large empirical differences in the predictability afforded by these different approaches, we are strongly in favor of the rational approach. It affords better theoretical or conceptual understanding of the domain of interest and may be best in terms of the stability of prediction over time. But, to our knowledge, convincing data on this question do not exist.

VALIDITY OF BIODATA

Various aspects of evidence concerning the validity of biodata have been collected. The findings reviewed next are arranged on the basis of the approach taken to justify the use of biodata measures in various situations.

Criterion-Related Validity

Biographical data are predictive of a range of criteria (Allworth & Hesketh, 2000). Biodata have been shown to predict the traditional criteria of job (estimated .35; Schmidt & Hunter, 1998) and training program performance (estimated .30; Schmidt & Hunter, 1998). Biodata measures were highly related to career and salary progression of managers in studies reported by the Standard Oil Company (1962, as cited by Owens, 1976) and others in the oil industry beginning in the 1960s. The generalization of such validities was reported by Rothstein et al. (1990). Various other criteria have been considered, including job satisfaction (e.g., Mumford & Owens, 1984; Sides, Feild, Giles, Holley, & Armenakis, 1984), turnover (e.g., Barrick & Zimmerman, 2005; Griffeth, Hom, & Gaertner, 2000), leadership

(e.g., Russell et al., 1990), and absenteeism (e.g., Schmitt et al., 2003). Validities for biodata measures vary across criteria and jobs, but corrected correlations tend to fall within the .30 to .40 range on average (Hunter & Hunter, 1984; Reilly & Chao, 1982; Schmitt, Gooding, Noe, & Kirsch, 1984). Breaugh (2009) pointed out that the size of correlations may depend on the type of validation design (concurrent, predictive) used, but biodata instruments have considerable validity in selection contexts regardless of the validation design used (Bliesener, 1996).

Incremental Validity

In studies of validity, researchers typically test how much incremental variance in a criterion is attributable to biodata above and beyond the variance accounted for by measures of general mental ability or personality. The latter are seen as fundamental selection tools beyond which biodata measures must prove useful (Mount, Witt, & Barrick, 2000) if a practitioner is to spend time developing these measures. Studies of biodata's incremental validity have supported its usefulness, although the extent to which it provides incremental predictive power has been shown to depend on the nature of the biodata and the construct being predicted. Allworth and Hesketh (2000) found that their job requirements biodata scale accounted for 6.5% of unique variance in ratings of employees' typical performance when it was added into a regression after cognitive ability. Both Dean and Russell (1998, as cited by Mount et al., 2000) and Dean (2004) noted the ability of their biographical data instrument to provide incremental validity for predicting training program performance beyond the validity shown by a test of general mental ability; an additional 9% of the variance in the performance criterion was explained by biodata in Dean's (2004) case. Biodata instruments have shown incremental validity over cognitive ability across performance criteria of varying specificity (i.e., specific performance domains, overall performance). Karas and West (1999) reported that their biodata scales explained 3.5% to 13.1% of unique variance in specific performance domains when various biodata composites were entered into hierarchical regressions after cognitive ability; 9% of unique variance in overall performance was explained by a combination of the biodata scales. Thus, when various aspects of performance are considered separately, the incremental validity of biodata over general mental ability may depend in part on the particular performance criterion. Allworth and Hesketh (1998, as cited by Mount et al., 2000), for instance, found that their change-oriented biodata scales accounted for the most unique variance over and above cognitive ability when the criterion was adaptive performance (vs. task or contextual performance).

One may try to turn to meta-analyses for an indication of biodata's incremental validity, but as Bliesener (1996) pointed out, meta-analyses on biographical data evaluate the validity of the method rather than the predictor constructs biodata instruments are designed to measure and thus do not present a clear picture of biodata's validity. In their meta-analysis of selection methods, Schmidt and Hunter (1998) found that the incremental validity of biographical data measures over general mental ability was .01 on average. They associated this small 2% increase in validity (from .51 to .52) with the high correlation between general mental ability and biographical data instruments (.50; Schmidt & Hunter, 1998), although this is certain to vary with the nature of the biodata items considered. Given the concern with meta-analyses of biographical data measures when the validities of instruments that measure different constructs are aggregated, superior conclusions should be derived from individual studies on biodata's validity.

Other individual studies have demonstrated the incremental validity of biodata over measures of personality (e.g., McManus & Kelly, 1999). Although researchers frequently consider incremental validity of biodata over cognitive ability or personality rather than both, some work has demonstrated biodata's incremental validity over a combination of cognitive ability and personality. For example, Mount et al. (2000) found that despite some overlap between their biodata scales and their personality and general mental ability constructs, the biodata scales did explain unique variance on three of four performance criteria.

However, performance is a typical but not the only criterion used in past research on biodata

validity. Other criteria have included attrition (Mael & Ashforth, 1995), absenteeism (Schmitt et al., 2003), and leadership (Mael & Hirsch, 1993). Thus, biodata measures have the potential to account for incremental variance over that explained by general mental ability or personality for criteria other than performance. Not surprisingly, though, the incremental validity of a biodata measure can be influenced by the method used to key biodata items (e.g., Mael & Hirsch, 1993) as well as the items themselves.

Construct and Content Validity

Evaluation of biodata measures has been characterized by a relative overreliance on criterion-related validity (Stokes & Cooper, 2001). Even though researchers recognize conceptual relations between biodata items and the criterion of interest as desirable for enhancing theoretical understanding of predictor–criterion relations, content and construct validity have received considerably less attention (Allworth & Hesketh, 2000), and research has provided relatively little explication of the psychological constructs underlying the observed relationships between biodata and various criteria (Mumford, Snell, & Reiter-Palmon, 1994), even though there have been repeated calls for greater attention to the nature of the constructs underlying biodata (e.g., Mumford, 1999). It is not surprising, then, that biodata measures are often criticized on the grounds of the scarcity of evidence for their content and construct validity (Mumford et al., 1996). Russell (1994) pointed out that a major deterrent to development of content- and construct-valid measures is that procedures for systematic creation of biodata items are lacking.

Some research has focused on the construct validity of biodata. Collins and Schmidt (1993) related biodata scores to a type of counterproductive work behavior. They used a biodata instrument to discriminate between employees who did and did not commit white-collar crimes and found that white-collar criminals received higher scores on two of their biodata scales: Extracurricular Activity and Social Extraversion. In another study, Mael (1995) used biodata to determine what individual differences contribute to swimming proficiency among White and Black cadets. Still other examples include work by Oswald et al. (2004), Schmitt et al. (1999), Schoenfeldt (1999), and Stokes et al. (1990). Mount et al. (2000), reviewing the literature, noted that biodata scales may measure various individual difference constructs, underscoring the fact that biodata represent a method of measurement rather than a construct per se. Mumford (1999) provided extensive suggestions on considerations for future research on construct validity of biodata measures. It is also the case that biodata items are often multiply determined or reflective of multiple constructs. Respondents may endorse an item or a specific option on a scale for different underlying reasons. For example, respondents to the second item in Table 25.1 may respond "0" partly as a function of the discretionary money they have; others may respond the same way out of an inherent stinginess.

Validity Generalization

The reality that biodata measures are typically custom made for organizations has traditionally been seen as limiting the use of biodata instruments by organizations for which they were not designed (Hunter & Hunter, 1984). Recent evidence to the contrary has emerged, however. To cite several examples, Brown's (1981) biodata scale proved to be a valid predictor of sales volume across 12 organizations, Rothstein et al.'s (1990) biodata scale predicted supervisor performance across 79 organizations, and Carlson, Scullen, Schmidt, Rothstein, and Erwin (1999) successfully used a biodata measure to predict promotion rates across 24 organizations. Research has indicated that biodata measures can be strategically designed to generalize across different jobs and organizations (e.g., Rothstein et al., 1990; Wilkinson, 1997). Several elements seem desirable if one is to expect generalizable results: large sample sizes, participation of multiple organizations (e.g., via a consortium), and cross-organizational keying of biodata scales (Rothstein et al., 1990). The relevance of criteria across organizations will be an important determinant of biodata's validity generalization in cases in which biodata items are chosen for the final instrument on the basis of their relationships with the criterion (Mount et al., 2000). The extent to

which biodata measures will be transportable will be determined by the amount of correspondence between the criterion used to develop the scale and the other criteria the biodata measure will be used to predict. Breaugh (2009) recommended designing biodata instruments with particular jobs in mind (e.g., supervisor) when transportability across organizations is desired. This recommendation simply underscores the need for an understanding of the constructs underlying the biodata measures and the fact that biodata can be developed to assess a variety of constructs.

Finally, evidence on the stability of biodata validity over time is mixed. Although some researchers have found evidence of temporal stability of biodata over considerable time periods (e.g., Brown, 1978; Carlson et al., 1999; Rothstein et al., 1990), others have not (e.g., Dunnette, Kirchner, Erickson, & Banas, 1960; Wernimont, 1962). It has been suggested that biodata measures should retain their criterion-related validity over a period of time to the extent that job positions and the target population remain fairly stable, changes in the environment (e.g., labor market, personnel policies) are minimal, and range restriction on the predictor or outcome does not reduce validity (Breaugh, 2009; Hogan, 1994). Maintaining the security of the scoring key over time is an important consideration in preventing validity from decaying (Brown, 1978).

Utility

Of primary concern in considering the usefulness of biodata in making selection decisions is its incremental validity over other instruments that are available; this issue is addressed in the Validity of Biodata section. Objectively scored biodata are easy to administer, and a large number of responses can be collected electronically or by paper and pencil in a relatively short period of time. However, there are two other important concerns. First, many biodata items are transparent as to their purpose and hence susceptible to applicants' attempts to inflate their status on job-related measures in any high-stakes situation. A second concern is whether the instruments are acceptable to candidates and organizations as well as consistent with legal constraints on what can and cannot be asked in an employment situation.

Social Desirability

Mean differences between applicants' and incumbents' responses to noncognitive measures similar to biodata are often substantial (e.g., Hough et al., 1990; Jackson, Wroblewski, & Ashton, 2000; Rosse, Stecher, Miller, & Levin, 1998). This difference is usually attributed to applicants' motivation to present themselves in the best possible light in high-stakes situations. Lautenschlager (1994) reviewed the results of 12 studies published between 1950 and 1990 that examined the presence and magnitude of the response distortion of biodata items. Distortion was examined in a variety of ways (e.g., responses to the same items on two occasions, correlations with external verifiable measures of the items, comparisons of test takers instructed to fake with those instructed to respond honestly, comparisons of applicants with others directed to respond honestly), so any attempt to examine the effect size associated with faking was not possible. Lautenschlager found standardized mean differences ranging from .2 to .54 across these 11 scales. Lautenschlager concluded that biodata instruments are fakeable and that in some instances people do fake them. He also concluded that there was some evidence that more objective and verifiable (see earlier discussion on item types) items were less susceptible to faking. In a more recent unpublished report, Fandre et al. (2008) compared the responses of a group of college student applicants ($n = 850$) with those of a group of already-admitted college students ($n = 2{,}756$) on a set of 11 rationally derived biodata measures and found differences of approximately 0.5 standard deviation, with applicants scoring higher.

Lautenschlager (1994) also indicated that direct warnings that responses would be verified served to produce lower scores on biodata instruments, with the implication being that distortion was reduced, a result that was subsequently supported by Dwight and Donovan (2003) using personality measures. However, because warnings could prove pointless if applicants have reason to doubt that verification will happen, instruments for which warnings are used should incorporate at least some clearly verifiable items. To effectively manage the impression of verifiability, more verifiable items could be put at the

beginning of the measure, and later verifiable items may be strategically intermixed with nonverifiable items (Mael, 1991). Schmitt and colleagues (Schmitt & Kunce, 2002; Schmitt et al., 2003) experimented with an alternative strategy for reducing faking: requiring that respondents elaborate on their biodata responses. For example, if respondents indicated they were involved in six volunteer organizations, they were asked to list the organizations (similarly, Items 1 and 2 in Exhibit 25.1 would lend themselves well to elaboration). Responses to the same items when elaboration was and was not requested differed by 0.5 to 0.9 standard deviation units with lower means in the elaborated conditions. However, effects for elaboration did not seem to carry over to responses to nonelaborated items (Schmitt et al., 2003). The latter result makes elaboration a less effective tool to reduce faking because elaboration of all items in an instrument would likely be impractical in most applied instances. In summary, potential users of biodata should be aware of the response distortion problem and attempt to minimize it by being strategic about types of items and instructions used.

Reactions to Biodata Items

To our knowledge, no studies exist of actual applicant reactions to the use of biodata in selection or admissions decisions (Breaugh, 2009). Nonapplicant reactions to biodata have, however, been considered at the overall instrument level by many researchers and with a wide variety of international samples (e.g., Anderson & Witvliet, 2008; Bertolino & Steiner, 2007; Ispas, Ilies, Iliescu, Johnson, & Harris, 2010; Marcus, 2003; Nikolaou & Judge, 2007; Schmitt, Oswald, Kim, Gillespie, & Ramsay, 2004; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993; Steiner & Gilliland, 1996). The research design has typically involved comparing reactions to different selection procedures (with biodata instruments included in that set) within samples (typically students or employees) and then comparing the rank ordering of instruments in terms of favorability across different samples. Although the location of biodata instruments in the list rank ordered by favorability varies depending on the sample considered, biodata instruments tend to fall somewhere in the middle of the list, with work samples and interviews fairly consistently ranking higher in favorability and graphology, integrity tests, and personal contacts ranking lower.

A few studies have also considered reactions to biodata items with particular attributes. Findings in these studies have indicated that items tend to be seen as more acceptable or less intrusive when they are verifiable, transparent, and ask for information that is public and regarding which inquiry is not questionable from a legal standpoint (Mael, Connerley, & Morath, 1996; Saks, Leck, & Saunders, 1995; Wallace, Page, & Lippstreu, 2006). Negative reactions to biodata have often been a result of the perception that items request information that is perceived to be an invasion of privacy and not relevant to the assessment of job-related skills. Legally, users of biodata must be careful that the use of their items does not represent a proxy for selection on the basis of membership in some protected group (i.e., a request to indicate where an applicant lives or went to school might be associated with ethnic status).

Reactions to biodata instruments and different types of biodata items seem to depend in part on respondents' characteristics (e.g., type of employment, demographic characteristics, cultural values; e.g., Mael et al., 1996; Ryan, Boyce, Ghumman, Jundt, & Schmidt, 2009). Mael et al. (1996), for instance, reported that their military sample was unique in that respondents saw negative items as more acceptable because of their ability to indicate about whether individuals would be able to psychologically and physically withstand pressure on the job. Because these respondents seemed to be using as a decision rule the question of whether responses to an item could provide an acceptable basis for excluding someone, reactions to more positive types of items were less favorable (e.g., that someone cannot report the same positive accomplishment as someone else is not going to matter for success in the military, so it should not be asked). Some evidence has shown that gender and race relate to perceptions of biodata items: women and Blacks may be more sensitive to potential insensitivity (because of invasiveness; Mael et al., 1996) and may have stronger privacy preferences (Connerley, Mael, & Morath, 1999). Finally, Ryan et al. (2009) found

some evidence to indicate that cultural values influence perceptions of biodata. Achievement orientation (seeing status as something that is earned via personal efforts and achievements as opposed to something given on the basis of birth, gender, money, etc.) and independence (seeing oneself as fairly independent of, as opposed to closely connected to, others) were positively related to perceptions of biodata's job relatedness in their sample of students drawn from across the globe.

Overall, the best recommendation with respect to consideration of reactions to biodata might be that items should be such that applicants and others perceive them to be job related, unfakeable, and not overly personal in nature (Ryan & Huth, 2008).

FUTURE RESEARCH AND USE

Many areas exist in which research might better inform the investigator who wants to use biodata to better understand or predict behavior. Four are highlighted. First, most of the research cited in this chapter has used biodata in a work or educational context, most often to predict work outcomes. Biodata measures should be useful in understanding social psychological, cognitive, and clinical phenomena. Mumford has repeatedly espoused their use in understanding human development (e.g., Mumford & Stokes, 1992). Some examples exist of the use of biodata in several other subdisciplines of psychology as well, but they are relatively rare. Examining life experiences can help explain individual differences in risk taking and responses to illness (Zinn, 2005), understand how experiences of living with mental illness are unique for different demographic groups (e.g., for Black women; Sosulski, Buchanan, & Donnell, 2010), evaluate and guide the treatment of psychiatric disorders (Sunnqvist, Persson, Lenntorp, & Traskman-Bendz, 2007), and contribute to psychologists' understanding of how illnesses and disorders (e.g., posttraumatic stress disorder; Osuch et al., 2001) progress and how individuals (e.g., those with schizophrenia) learn to detect their early signs of relapse (C. Baker, 1995). Biodata can help elucidate individual differences in language development (Elardo, Bradley, & Caldwell, 1977), reading skills (Sénéchal & LeFevre, 2002), artistic achievement (Crozier, 2003), and musical achievement (Manturzewska, 1990) and in educational outcomes (Murasko, 2007). Examining what researchers in other disciplines sometimes call "life histories" can also help identify ways in which professional development and training could be improved for various occupations (e.g., for music service teachers; D. Baker, 2006).

Second, even though previous reviewers have consistently called for more research on the constructs underlying biodata, there remains very little consensus on this issue. Biodata represents a method of measurement, but the constructs indexed best using biodata remain to be identified. One thing that impedes progress here is often the emphasis on their use in a particular organizational context. It is certainly understandable that an organization interested in predicting job performance in some of its jobs will write items that will reflect background data and experiences that are reflective of its jobs. In this context, it might be good to remember the distinction between worker-oriented items (that are likely to be related to psychological constructs) and job-oriented items (that directly index experience on tasks associated with a particular job). The latter are likely to be predictive of performance on a given job, but the former are likely to provide a greater understanding of the nature of the traits associated with performance in a broader array of contexts. Several examples of research that focused on constructs were cited in this chapter; these models should be more consistently used.

Owens's (1976) work on using biodata clusters to profile student capabilities still holds promise. He and his colleagues provided impressive evidence that the lives and careers of people with different biographical profiles followed very different and predictable paths. This work has not often been pursued by other researchers and should provide valuable insight into people's career trajectories.

Third, faking remains a significant problem with the use of biodata, probably more so now than in the past because many biodata items now used ask for an opinion, preference, or value. The answer to these items is rarely objective or verifiable and, as with personality items, the answer that would portray a respondent in the most favorable light is

fairly obvious. Hence, in any high-stakes situation, responses are likely to be inflated and not represent respondents' true standing on the targeted construct.

Finally, given the considerable implications applicant reactions have for organizations (Hausknecht, Day, & Thomas, 2004), this is an area in which more work should be done (even if it is only with nonapplicant groups). Research on reactions to particular types of biodata items is currently fairly scarce. More than 20 years ago, Mael (1991) called for more research to investigate the practical benefits of choosing to use biodata items with particular attributes. With respect to reactions, only a few researchers have answered this call, and one of them is Mael himself. Given legal considerations, little is left to item writers' discretion with regard to how intrusive biodata items can get (Smither et al., 1993), but more is left to their discretion with regard to some other item attributes (e.g., transparent vs. subtle). Considering the fact that the set of life experiences that could be relevant is larger than the set of experiences that would be easily recognized as relevant (i.e., face valid; Mael, 1991) and the expectation that individuals will react more negatively to subtle items, researchers might consider what strategies could be used (e.g., framing of the questions asked) to aid in the ability to ask the questions considered predictive while not negatively affecting applicant reactions. Also, more could be learned about what individual differences contribute to reactions and how reactions may differ on the basis of the medium used to administer biodata items (e.g., paper and pencil, computer, interactive voice response; Van Iddekinge et al., 2003). The latter is a particularly interesting issue given the increasing use of technology in selection (Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000) and would supplement existing research examining the comparability of biodata forms administered in different formats (paper and pencil vs. web based; Ployhart, Weekley, Holtz, & Kemp, 2003).

References

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York, NY: Holt, Rinehart & Winston.

Allworth, E., & Hesketh, B. (2000). Job requirements biodata as a predictor of performance in customer service roles. *International Journal of Selection and Assessment, 8*, 137–147. doi:10.1111/1468-2389.00142

Allworth, E. G., & Hesketh, B. (1998, April). *Generalizability of construct-oriented biodata scales in predicting adaptive performance.* Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, Texas.

Anderson, N., & Witvliet, C. (2008). Fairness reactions to personnel selection methods: An international comparison between the Netherlands, the United States, France, Spain, Portugal, and Singapore. *International Journal of Selection and Assessment, 16*, 1–13. doi:10.1111/j.1468-2389.2008.00404.x

Baker, C. (1995). The development of the self-care ability to detect early signs of relapse among individuals who have schizophrenia. *Archives of Psychiatric Nursing, 9*, 261–268. doi:10.1016/S0883-9417(95)80045-X

Baker, D. (2006). Life histories of music service teachers: The past in inductees' present. *British Journal of Music Education, 23*, 39–50. doi:10.1017/S026505170500673X

Barrick, M. R., & Zimmerman, R. D. (2005). Reducing voluntary, avoidable turnover through selection. *Journal of Applied Psychology, 90*, 159–166. doi:10.1037/0021-9010.90.1.159

Becker, T. E., & Colquitt, A. L. (1992). Potential versus actual faking of a biodata form: An analysis along several dimensions of item type. *Personnel Psychology, 45*, 389–406. doi:10.1111/j.1744-6570.1992.tb00855.x

Bertolino, M., & Steiner, D. D. (2007). Fairness reactions to selection methods: An Italian study. *International Journal of Selection and Assessment, 15*, 197–205. doi:10.1111/j.1468-2389.2007.00381.x

Bliesener, T. (1996). Methodological moderators in validating biographical data in personnel selection. *Journal of Occupational and Organizational Psychology, 69*, 107–120. doi:10.1111/j.2044-8325.1996.tb00603.x

Breaugh, J. A. (2009). The use of biodata for employee selection: Past research and future directions. *Human Resource Management Review, 19*, 219–231. doi:10.1016/j.hrmr.2009.02.003

Brown, S. H. (1978). Long-term validity of a personal history item scoring procedure. *Journal of Applied Psychology, 63*, 673–676. doi:10.1037/0021-9010.63.6.673

Brown, S. H. (1981). Validity generalization and situational moderation in the life insurance industry. *Journal of Applied Psychology, 66*, 664–670. doi:10.1037/0021-9010.66.6.664

Burisch, M. (1984). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18*, 81–98. doi:10.1016/0092-6566(84)90040-0

Carlson, K. D., Scullen, S. E., Schmidt, F. L., Rothstein, H., & Erwin, F. (1999). Generalizable biographical data validity can be achieved without multi-organizational development and keying. *Personnel Psychology, 52*, 731–755. doi:10.1111/j.1744-6570.1999.tb00179.x

Colangelo, N., Kerr, B., Hallowell, K., Huesman, R., & Gaeth, J. (1992). The Iowa Inventiveness Inventory: Toward a measure of mechanical inventiveness. *Creativity Research Journal, 5*, 157–163. doi:10.1080/10400419209534429

Collins, J. M., & Schmidt, F. L. (1993). Personality, integrity, and white collar crime: A construct validity study. *Personnel Psychology, 46*, 295–311. doi:10.1111/j.1744-6570.1993.tb00875.x

Connerley, M. L., Mael, F. A., & Morath, R. A. (1999). "Don't ask—Please tell": Selection privacy from two perspectives. *Journal of Occupational and Organizational Psychology, 72*, 405–422. doi:10.1348/096317999166752

Crafts, J. (1991). *Using biodata as a selection instrument*. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation. (ERIC Document Reproduction Service No. ED 338702)

Crozier, W. R. (2003). Individual differences in artistic achievement: A within-family case study. *Creativity Research Journal, 15*, 311–319. doi:10.1207/S15326934CRJ1504_1

Dean, M. A. (2004). An assessment of biodata predictive validity across multiple performance criteria. *Applied H. R. M. Research, 9*, 1–12.

Dean, M. A., & Russell, C. J. (1998, April). *A comparison of and biodata criterion-related validity in a sample of air traffic controllers.* Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Dean, M. A., Russell, C. J., & Muchinsky, P. M. (1999). Life experiences and performance prediction: Toward a theory of biodata. In G. Ferris (Ed.), *Research in personnel and human resources management* (pp. 245–281). Greenwich, CT: JAI Press.

Dunnette, M. D., Kirchner, W. K., Erickson, J., & Banas, P. (1960). Predicting turnover among female office workers. *Personnel Administration, 23*, 45–50.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23. doi:10.1207/S15327043HUP1601_1

Elardo, R., Bradley, R., & Caldwell, B. M. (1977). A longitudinal study of the relation of infants' home environments to language development at age three. *Child Development, 48*, 595–603. doi:10.2307/1128658

England, G. W. (1971). *Development and use of weighted application blanks* (Rev. ed.). Minneapolis: University of Minnesota, Industrial Relations Center.

Fandre, J., Pleskac, T. J., Quinn, A., Schmitt, N., Sinha, R., & Zorzie, M. (2008). *Comparisons of the responses of college applicants and college students to biodata and situational judgment measures* (College Board Report). East Lansing: Michigan State University.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327–358. doi:10.1037/h0061470

Galton, F. (1902). *Life history album* (2nd ed.). New York, NY: McMillan.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* New York, NY: Wiley.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26*, 463–488. doi:10.1177/014920630002600305

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639–683. doi:10.1111/j.1744-6570.2004.00003.x

Henry, E. R. (1966). *Research conference on the use of autobiographical data as psychological predictors.* Greensboro, NC: Creativity Research Institute & Richardson Foundation.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced choice normative measures. *Psychological Bulletin, 74*, 167–184. doi:10.1037/h0029780

Hogan, J. B. (1994). Empirical keying of background data measures. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 69–107). Palo Alto, CA: Consulting Psychologists Press.

Hough, L., & Paullin, C. (1994). Construct-oriented scale construction: The rational approach. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research and use of biographical information in selection and performance prediction* (pp. 109–146). Palo Alto, CA: Consulting Psychologists Press.

Hough, L. M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology, 69*, 135–146. doi:10.1037/0021-9010.69.1.135

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581–595. doi:10.1037/0021-9010.75.5.581

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98. doi:10.1037/0033-2909.96.1.72

Ispas, D., Ilies, A., Iliescu, D., Johnson, R. E., & Harris, M. M. (2010). Fairness reactions to selection methods: A Romanian study. *International Journal of Selection and Assessment, 18*, 102–110. doi:10.1111/j.1468-2389.2010.00492.x

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371–388. doi:10.1207/S15327043HUP1304_3

Kadden, R. M., Litt, M. D., Donovan, D., & Cooney, N. L. (1996). Psychometric properties of the California Psychological Inventory Socialization scale in treatment-seeking alcoholics. *Psychology of Addictive Behaviors, 10*, 131–146. doi:10.1037/0893-164X.10.3.131

Karas, M., & West, J. (1999). Construct-oriented biodata development for selection to a differentiated performance domain. *International Journal of Selection and Assessment, 7*, 86–96. doi:10.1111/1468-2389.00109

Kilcullen, R. N., White, L. A., Mumford, M. D., & Mack, H. (1995). Assessing the construct validity of rational biodata scales. *Military Psychology, 7*, 17–28. doi:10.1207/s15327876mp0701_2

Lautenschlager, G. J. (1994). Accuracy and faking of background data. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook* (pp. 394–419). Palo Alto, CA: Consulting Psychologists Press.

Lefkowitz, J., Gebbia, M. I., Balsam, T., & Dunn, L. (1999). Dimensions of biodata items and their relationships to item validity. *Journal of Occupational and Organizational Psychology, 72*, 331–350. doi:10.1348/096317999166716

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44*, 763–792. doi:10.1111/j.1744-6570.1991.tb00698.x

Mael, F. A. (1995). Staying afloat: Within-group swimming proficiency for Whites and Blacks. *Journal of Applied Psychology, 80*, 479–490. doi:10.1037/0021-9010.80.4.479

Mael, F. A., & Ashforth, B. E. (1995). Loyal from day one: Biodata, organizational identification, and turnover among newcomers. *Personnel Psychology, 48*, 309–333. doi:10.1111/j.1744-6570.1995.tb01759.x

Mael, F. A., Connerley, M., & Morath, R. A. (1996). None of your business: Parameters of biodata invasiveness. *Personnel Psychology, 49*, 613–650. doi:10.1111/j.1744-6570.1996.tb01587.x

Mael, F. A., & Hirsch, A. C. (1993). Rainforest empiricism and quasi-rationality: Two approaches to objective biodata. *Personnel Psychology, 46*, 719–738. doi:10.1111/j.1744-6570.1993.tb01566.x

Manturzewska, M. (1990). A biographical study of the life-span development of professional musicians. *Psychology of Music, 18*, 112–139. doi:10.1177/0305735690182002

Marcus, B. (2003). Attitudes towards personnel selection methods: A partial replication and extension in a German sample. *Applied Psychology, 52*, 515–532. doi:10.1111/1464-0597.00149

McIntyre, T. M., Mauger, P. A., Margalit, B., & Figueiredo, E. (1989). The generalizability of aggressiveness and assertiveness factors: A cross-cultural analysis. *Personality and Individual Differences, 10*, 385–389. doi:10.1016/0191-8869(89)90003-2

McManus, M. A., & Kelly, M. L. (1999). Personality measures and biodata: Evidence regarding their incremental predictive validity in the life insurance industry. *Personnel Psychology, 52*, 137–148. doi:10.1111/j.1744-6570.1999.tb01817.x

Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology, 67*, 411–418. doi:10.1037/0021-9010.67.4.411

Mount, M. K., Witt, L. A., & Barrick, M. R. (2000). Incremental validity of empirically keyed biodata scales over GMA and the five factor personality constructs. *Personnel Psychology, 53*, 299–323. doi:10.1111/j.1744-6570.2000.tb00203.x

Mumford, M. D. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 61–138). Palo Alto, CA: Consulting Psychologists Press.

Mumford, M. D. (1999). Construct validity and background data: Issues, abuses and future directions. *Human Resource Management Review, 9*, 117–145. doi:10.1016/S1053-4822(99)00015-7

Mumford, M. D., Connelly, M. S., & Clifton, T. C. (1990). *The meaning of life history measures: Implications for scaling strategies*. Unpublished report, Center for Behavioral and Cognitive Studies, George Mason University.

Mumford, M. D., Costanza, D. P., Connelly, M. S., & Johnson, J. F. (1996). Item generation procedures

and background data scales: Implications for construct and criterion-related validity. *Personnel Psychology, 49*, 361–398. doi:10.1111/j.1744-6570.1996.tb01804.x

Mumford, M. D., & Owens, W. A. (1984). Individuality in a developmental context: Some empirical and theoretical considerations. *Human Development, 27*, 84–108.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 1–31. doi:10.1177/014662168701100101

Mumford, M. D., Snell, A. F., & Reiter-Palmon, R. (1994). Background data and development: Structural issues in the application of life history measures. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 555–581). Palo Alto, CA: Consulting Psychologists Press.

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 3, pp. 61–138). Palo Alto, CA: Consulting Psychologists Press.

Murasko, J. E. (2007). A lifecourse study on education and health: The relationship between childhood psychosocial resources and outcomes in adolescence and young adulthood. *Social Science Research, 36*, 1348–1370. doi:10.1016/j.ssresearch.2007.01.001

Nickels, B. J. (1994). The nature of biodata. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 1–16). Palo Alto, CA: Consulting Psychologists Press.

Nikolaou, I., & Judge, T. A. (2007). Fairness reactions to personnel selection techniques in Greece: The role of core self-evaluations. *International Journal of Selection and Assessment, 15*, 206–219. doi:10.1111/j.1468-2389.2007.00382.x

Osuch, E. A., Brotman, M. A., Podell, D., Geraci, M., Touzeau, P. L., Leverich, G. S., . . . Post, R. M. (2001). Prospective and retrospective life-charting in post-traumatic stress disorder (the PTSD-LCM): A pilot study. *Journal of Traumatic Stress, 14*, 229–239. doi:10.1023/A:1007860204298

Oswald, F. L., Schmitt, N., Kim, B. H., Gillespie, M. A., & Ramsay, L. J. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*, 187–207. doi:10.1037/0021-9010.89.2.187

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 609–644). Chicago, IL: Rand-McNally.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. *Journal of Applied Psychology, 64*, 569–607. doi:10.1037/0021-9010.64.5.569

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56*, 733–752. doi:10.1111/j.1744-6570.2003.tb00757.x

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1–62. doi:10.1111/j.1744-6570.1982.tb02184.x

Reiter-Palmon, R., DeFilippo, B., & Mumford, M. D. (1990, March). *Differential predictive validity of positive and negative response options to biodata items*. Paper presented at the annual meeting of the Southeastern Psychological Association, Atlanta, GA.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880–887. doi:10.1037/0021-9010.85.6.880

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644. doi:10.1037/0021-9010.83.4.634

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75*, 175–184. doi:10.1037/0021-9010.75.2.175

Russell, C. J. (1994). Generation procedures for biodata items: A point of departure. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 17–38). Palo Alto, CA: Consulting Psychologists Press.

Russell, C. J., Matson, J., Devlin, S. E., & Atwater, D. (1990). Predictive validity of biodata items generated from retrospective life experience essays. *Journal of Applied Psychology, 75*, 569–580. doi:10.1037/0021-9010.75.5.569

Ryan, A. M., Boyce, A. S., Ghumman, S., Jundt, D., & Schmidt, G. (2009). Going global: Cultural values and perceptions of selection procedures. *Applied Psychology, 58*, 520–556. doi:10.1111/j.1464-0597.2008.00363.x

Ryan, A. M., & Huth, M. (2008). Not much more than platitudes? A critical look at the utility of applicant reactions research. *Human Resource Management Review, 18*, 119–132. doi:10.1016/j.hrmr.2008.07.004

Saks, A. M., Leck, J. D., & Saunders, D. M. (1995). Effects of application blanks and employment equity on applicant reactions and job pursuit intentions. *Journal of Organizational Behavior, 16*, 415–430. doi:10.1002/job.4030160504

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274. doi:10.1037/0033-2909.124.2.262

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422. doi:10.1111/j.1744-6570.1984.tb00519.x

Schmitt, N., Jennings, D., & Toney, R. (1999). Can we develop biodata measures of hypothetical constructs? *Human Resource Management Review, 9*, 169–183. doi:10.1016/S1053-4822(99)00017-0

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of four-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology, 94*, 1479–1497. doi:10.1037/a0016810

Schmitt, N., & Kunce, C. (2002). The effects of required elaboration of answers to biodata questions. *Personnel Psychology, 55*, 569–587. doi:10.1111/j.1744-6570.2002.tb00121.x

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., & Ramsay, L. J. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88*, 979–988. doi:10.1037/0021-9010.88.6.979

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., & Ramsay, L. J. (2004). The impact of justice and self-serving bias explanations for the perceived fairness of different types of selection tests in college admissions. *International Journal of Selection and Assessment, 12*, 160–171. doi:10.1111/j.0965-075X.2004.00271.x

Schoenfeldt, L. F. (1999). From dust bowl empiricism to rational constructs in biographical data. *Human Resource Management Review, 9*, 147–167. doi:10.1016/S1053-4822(99)00016-9

Sénéchal, M., & LeFevre, J.-A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development, 73*, 445–460. doi:10.1111/1467-8624.00417

Sides, E. H., Feild, H. S., Giles, W. F., Holley, W. H., & Armenakis, A. A. (1984). Biographical data: A neglected tool for career counselling. *Human Resource Planning, 7*, 151–156.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49–76. doi:10.1111/j.1744-6570.1993.tb00867.x

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Toward a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9*, 219–242. doi:10.1016/S1053-4822(99)00019-4

Sosulski, M. R., Buchanan, N. T., & Donnell, C. M. (2010). Life history and narrative analysis: Feminist methodologies contextualizing Black women's experiences with severe mental illness. *Journal of Sociology and Social Welfare, 37*, 29–57.

Standage, K., Smith, D., & Norman, R. (1988). A classification of respondents to the CPI Socialization Scale: Associations with psychiatric diagnosis and implications for research. *Personality and Individual Differences, 9*, 231–236. doi:10.1016/0191-8869(88)90084-0

Steiner, D. D., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology, 81*, 134–141. doi:10.1037/0021-9010.81.2.134

Steiner, D. D., & Gilliland, S. W. (2001). Procedural justice in personnel selection: International and cross-cultural perspectives. *International Journal of Selection and Assessment, 9*, 124–137. doi:10.1111/1468-2389.00169

Stokes, G. S., & Cooper, L. A. (2001). Content/construct approaches in life history form development for selection. *International Journal of Selection and Assessment, 9*, 138–151. doi:10.1111/1468-2389.00170

Stokes, G. S., Toth, C. S., Searcy, C. A., Stroupe, J. P., & Carter, G. (1999). Construct/rational biodata dimensions to predict salesperson performance: Report on the U.S. Department of Labor sales study. *Human Resource Management Review, 9*, 185–218. doi:10.1016/S1053-4822(99)00018-2

Sunnqvist, C., Persson, U., Lenntorp, B., & Traskman-Bendz, L. (2007). Time geography: A model for psychiatric life charting? *Journal of Psychiatric and Mental Health Nursing, 14*, 250–257. doi:10.1111/j.1365-2850.2007.01071.x

Van Iddekinge, C. H., Eidson, C. E., Jr., Kudisch, J. D., & Goldblatt, A. M. (2003). A biodata inventory administered via interactive voice response (IVR) technology: Predictive validity, utility, and subgroup differences. *Journal of Business and Psychology, 18*, 145–156. doi:10.1023/A:1027340913460

Wallace, J. C., Page, E. E., & Lippstreu, M. (2006). Applicant reactions to pre-employment application blanks: A legal and procedural justice perspective. *Journal of Business and Psychology, 20*, 467–488. doi:10.1007/s10869-005-9007-0

Wernimont, P. F. (1962). Reevaluation of a weighted application blank for office personnel. *Journal of Applied Psychology, 46*, 417–419. doi:10.1037/h0043645

Whitney, D. J., & Schmitt, N. (1997). Relationship between culture and responses to biodata employment items. *Journal of Applied Psychology, 82*, 113–129. doi:10.1037/0021-9010.82.1.113

Wilkinson, L. J. (1997). Generalizable biodata? An application to the vocational interests of managers. *Journal of Occupational and Organizational Psychology, 70*, 49–60. doi:10.1111/j.2044-8325.1997.tb00630.x

Zinn, J. O. (2005). The biographical approach: A better way to understand behaviour in health and illness. *Health, Risk, and Society, 7*, 1–9. doi:10.1080/13698570500042348

CHAPTER 26

ASSESSMENT OF LEADERSHIP

Nancy T. Tippins

There is little doubt in modern businesses that leadership is a requirement for an organization's success. Without leadership, there is no direction for today and no vision of the future. Despite a widespread belief that the need for leadership is critical, psychologists' understanding of exactly what it entails is highly varied. Yet, to assess leadership, psychologists working in business and industry must first understand what it is and what to measure. Unfortunately, there are many definitions of *leadership,* which makes defining appropriate measures difficult. To focus the discussion, this chapter defines *leadership* as the accomplishment of work through others and operationalizes the concept in terms of the knowledge, skills, abilities, and other characteristics (KSAOs) required to perform well in leadership roles, Thus, with this definition, leadership can occur at any level in the organization and can include both formal and informal leadership roles. However, because assessment for informal leadership positions is rarely done, this chapter concentrates on formal, assigned leadership positions, in other words, supervisory, managerial, and executive roles. *Supervisory roles* are defined as those in which the incumbent supervises others who perform work but do not supervise others. The direct reports of managers are people who manage other people, either supervisors or other managers. Executives also manage other managers but do so from the top of the organization. Silzer (2002) suggested that executives are generally considered to include general managers, corporate officers, and heads of major organizational functions and business units and senior executives to include corporate officers, executive committee members, and chief executive officers. (In this chapter, executives and senior executives are not distinguished.) In addition to supervisory responsibilities, each higher level of leadership generally has a broader scope of responsibility and authority.

Although discussing each category of leadership (supervisory, managerial, and executive) as a discrete group would be useful, the research is not so clean. The lines between supervisors and managers and between managers and executives are often blurred. Research reports and journal articles frequently classify participants as managers and fail to say at what level of the organization these people manage.

The two primary purposes for leadership assessment are (a) selection and (b) development. (Promotion is considered a special case of selection in which only internal candidates are considered for higher level positions that involve a greater scope of responsibility.) Some assessments are designed for one or the other purpose, but many are designed with both purposes in mind. Frequently, candidates who are evaluated for selection or promotion purposes also receive assessment feedback designed to guide their developmental activities. Organizations that sponsor developmental assessments often have access to results that may directly or indirectly influence future placement and promotion decisions. Both purposes are considered in the discussion of assessment techniques.

This chapter begins with a discussion of the foundations for assessment, job analysis, and

DOI: 10.1037/14047-026
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

competency models and then summarizes the research on major forms of leadership assessment tools: cognitive ability, personality, biodata, situational judgment tests (SJTs), 360-degree feedback instruments, individual assessment, and assessment centers. The research reported includes summaries of the validities based on meta-analyses. When subgroup differences have been discussed in other chapters in this volume, the reader is referred to those chapters. Otherwise, brief summaries of typical findings are provided. For each assessment type, the advantages and disadvantages for practice are briefly reviewed.

Local validation studies are difficult to conduct for many leadership positions, especially for higher level positions when the population is small. Often, there are simply not enough people who are assessed and placed into an executive position. In other cases, the reasons are administrative. Many at higher levels of the organization are unwilling to participate either by being evaluated or by providing performance information about others. Although the specific reasons for nonparticipation are unknown, typical responses to participation requests refer to limits on the time for or lack of commitment to such an endeavor. A few of those asked to participate are probably afraid to display their weaknesses. Additionally, the problem of obtaining an adequate criterion can derail a validity study and limit the observed relationship between assessment results and on-the-job performance in leadership roles.

Except for cognitive abilities, most meta-analyses have not separated leadership roles from other kinds of jobs. The reason for this lack of meta-analyses is technical—there are not a sufficient number of criterion-related validity studies on leaders alone. Most of the meta-analyses that are available are based on studies of many diverse jobs. Few have looked at jobs that could be classified as leadership positions, and fewer still have looked at the moderating effects of variables related to leadership positions. Job complexity and its effect on the validities for cognitive ability is the primary exception.

Despite the dearth of evidence for the validity of various assessment instruments used in identifying individual strengths and weaknesses and predicting performance in leadership roles, it should be acknowledged that content-oriented approaches to accumulating evidence of validity are an acceptable approach when "the content of the selection procedure is representative of important aspects of performance on the job for which the candidates are to be evaluated" (Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice, 1978, § 5.B). Often, small sample sizes rule out a criterion-related validity study. Thus, many assessments consisting of tasks that mimic actual work (e.g., simulations, role plays, in baskets) are validated locally using a content-oriented approach. Similarly, assessment procedures that ask for descriptions of certain kinds of job-relevant behaviors, such as interviews that ask the candidate to describe his or her own actions or 360-degree feedback instruments that ask the candidate and others to rate certain behaviors, are frequently validated using a content-oriented strategy to ensure job relevancy. Although anecdotal evidence has suggested widespread use of content-oriented approaches to establishing validity, these types of studies are rarely published, and no accepted procedure exists for systematically synthesizing them even if they were available. Nevertheless, for the practitioner, a careful study of the job requirements and a systematic comparison of the constructs measured by the assessment tool with those job requirements may establish sufficient evidence of the validity of the inferences to be made, particularly when the sample size is very small.

Regardless of whether one believes that validation is or is not required or that it is wise or unwise not to collect validity evidence locally, it should be acknowledged that evidence of validity is not always established in a systematic way. Many instruments are by design very closely related to leadership behaviors. For example, 360-degree evaluations or work simulations often reflect the work of a leader in a particular organization. Frequently, the practitioner as well as his or her colleagues within the organization see little need to ensure job relevance through a series of content validation linkages. Moreover, work activities and knowledge, skills, abilities, and other characteristics (KSAOs) for a specific group of leader jobs are often not available, and

the challenges of job analysis with executives can be significant. In addition, the need for evidence of validity collected locally may not be recognized either because the assessment process is for development purposes and is not intended for purposes of employee selection or because the risk of a challenge from candidates for jobs at high levels is perceived to be low.

JOB ANALYSIS FOR LEADERSHIP ASSESSMENT

The basis for ensuring that an assessment process is job relevant and measures the right things is job analysis. The methodology for job analysis has been well documented in other sources (see Pearlman & Sanchez, 2010, for a recent summary, as well as Chapter 23, this volume) and is not the goal of this chapter; instead, the intent here is to provide a brief synopsis of previous efforts to define the KSAOs required of leaders.

O*NET Data

The discussion of requirements of leadership positions begins with O*NET, the U.S. Department of Labor's database of occupational information, because it is a starting point for many practitioners who develop or use leadership assessments (see http://www.onetonline.org). Although most practitioners are highly attuned to the unique requirements of managerial jobs in a particular organization, they are also aware of the commonalities that usually exist across organizations. In addition, the information in the O*NET reflects the artificiality of the strict classification of jobs into supervisory, managerial, or executives roles and highlights the need for careful consideration of the specific job requirements.

The level of specificity provided for supervisory occupations in the O*NET makes the identification of common work activities or KSAOs across similar positions difficult. For example, in 2010, 26 occupations were listed with *supervisor* in the title. In addition, supervisors and managers are not distinguished in many occupational fields. For example, many titles read "first-line supervisor/manager of. . . ." The differences between the work activities and skill requirements of managers and supervisors are not clear; however, the inclusion of at least 81 occupations with the words *manager* or *management* indicates some implicit differences between them and the 26 occupations with *supervisor* in the title. To add to the confusion, *chief executives* are the sixth supervisory occupation listed.

In a study of O*NET data for supervisors, Tippins (2009a) found no core set of work activities or skill requirements among a set of 19 occupational titles containing the word *supervisor*. Table 26.1 shows the most common work activities across 19 supervisory jobs. Not one of the O*NET's general work activities was endorsed in all supervisory roles. As important as what was endorsed is what was not endorsed. Work activities that are often expected of supervisors such as coaching and developing others, developing and building teams, scheduling work and activities, and training and teaching others were required in fewer than half of these positions. Similar patterns emerged for skills (see Table 26.2). Not all skills were required of all supervisory jobs; however, four abilities were common to all 19 jobs: oral expression, oral comprehension, deductive reasoning, and problem sensitivity (see Table 26.3).

In addition to the lack of differentiation between supervisor and manager, the managerial occupations also seem to overlap with executive positions. For example, the representative job titles for the occupation general and operations manager include

TABLE 26.1

Work Activities and Number of Jobs out of 19 Supervisory Jobs

Work activity	No. of jobs
Making decisions and solving problems	18
Communicating with supervisors, peers, or subordinates	16
Getting information	17
Organizing, planning, and prioritizing work	11
Coordinating the work and activities of others	11
Identifying objects, actions, and events	11
Establishing and maintaining interpersonal relationships	10
Guiding, directing, and motivating subordinates	10
Resolving conflicts and negotiating with others	10

TABLE 26.2

Skills for and Number of Jobs for 19 Front-Line Supervisory Jobs

Skills	No. of jobs
Active listening	18
Critical thinking	18
Time management	16
Reading comprehension	15
Instructing	14
Monitoring	14
Speaking	13
Management of personnel resources	13
Judgment and decision making	12

TABLE 26.3

Abilities for and Number of Jobs for 19 Front-Line Supervisory Jobs

Abilities	No. of jobs
Oral expression	19
Oral comprehension	19
Deductive reasoning	19
Problem sensitivity	19
Inductive reasoning	17
Speech clarity	16
Speech recognition	15
Near vision	13
Written comprehension	13

operations manager, general manager, director of operations, plant manager, store manager, facilities manager, plant superintendent, vice president of operations, warehouse manager, and chief operating officer. Some of these titles appear to fall in the higher level executive category as described by Silzer (2002). Although executive positions appear to be scattered throughout the list of manager occupations, only one occupation, chief executive, has *executive* in its title. Exhibit 26.1 summarizes those task requirements. Despite these concerns about how the O*NET data are organized, the database provides a rich source of information about jobs that might be generally classified as managerial and offers a beginning point for local research.

Exhibit 26.1
O*NET Tasks for Chief Executive Occupation

- Direct and coordinate an organization's financial and budget activities to fund operations, maximize investments, and increase efficiency.
- Confer with board members, organization officials, and staff members to discuss issues, coordinate activities, and resolve problems.
- Analyze operations to evaluate performance of a company and its staff in meeting objectives and to determine areas of potential cost reduction, program improvement, or policy change.
- Direct, plan, and implement policies, objectives, and activities of organizations or businesses to ensure continuing operations, to maximize returns on investments, and to increase productivity.
- Prepare budgets for approval, including those for funding and implementation of programs.
- Direct and coordinate activities of businesses or departments concerned with production, pricing, sales, or distribution of products.
- Negotiate or approve contracts and agreements with suppliers, distributors, federal and state agencies, and other organizational entities.
- Review reports submitted by staff members to recommend approval or to suggest changes.
- Appoint department heads or managers and assign or delegate responsibilities to them.
- Direct human resources activities, including the approval of human resources plans and activities, the selection of directors and other high-level staff, and establishment and organization of major departments.

Note. Retrieved from http://www.onetonline.org/link/summary/11-1011.00

Literature on Job Requirements for Leaders

Finding comprehensive studies of both the task and KSAO requirements of a broad spectrum of leadership jobs is difficult (in contrast to studies of only time spent or studies of managerial jobs or supervisory jobs). Documentation of those studies that are done is often proprietary and rarely published. In addition, the research literature on task and KSAO requirements for leadership roles is often vague on exactly what jobs are included in a particular study. Instead, many of the job analyses simply state that manager jobs were studied. Nevertheless, some examples of research on the requirements of leadership positions are provided.

Early research on managerial jobs measured the amount of time that managers spent on various tasks. Mahoney, Jerdee, and Carroll (1965) investigated how 452 managers in 13 companies spent their time and defined eight management functions: planning, investigating, coordinating, evaluating, supervising, staffing, negotiating, and representing. Mintzberg (1975) also investigated the amount of time spent in various activities. Although managers do have some regular activities, he found that managers spent a great deal of time responding to unplanned demands, noting that their activities were characterized by "brevity, variety, and discontinuity" (Mintzberg, 1975, p. 50). Hemphill (1960) described 10 dimensions of leadership of managerial jobs. When this work was updated by Tornow and Pinto in 1976, 13 dimensions emerged, with only some of them remaining the same as the earlier study. More recently, Schippmann, Prien, and Hughes (1991) reviewed 21 job analyses that identified the tasks associated with managerial jobs and 10 analyses that investigated job skills and qualitatively developed 21 task dimensions and 22 job skill dimensions. Mumford, Campion, and Morgeson (2007) defined four broad categories of management skills: cognitive skills, interpersonal skills, business skills, and strategic skills.

Erker, Cosentino, and Tamanini (2010) emphasized the changing demands on leaders in modern organizations and compared the behavioral requirements for leadership positions in Fortune 1,000 companies before and after 2000. Several hundred job analyses were reviewed, and the 11 most common competencies critical to success before and after 2000 were identified on the basis of the percentage of time the competency was identified as important. Although the communications and decision-making competencies remained the two most important in both time periods, Erker et al. concluded that before 2000, managers were more technically oriented, and after 2000, managers balanced technical competence with interpersonal skills.

R. Hogan and Warrenfeltz (2003) identified four classes of behavioral competence:

- leadership skills—building and motivating a high-performing team;
- intrapersonal skills—regulating one's emotions, easily accommodating to authority;
- interpersonal skills—building and maintaining relationships; and
- business skills—planning, budgeting, coordinating, and monitoring business activities.

Subsequently, the Hogan researchers (Davies, Hogan, Foster, & Elizondo, 2005; J. Hogan, Davies, & Hogan, 2007) linked each of these to one of the factors in the five factor model of personality, identifying the links between established personality measures and leadership behaviors (intrapersonal and conscientiousness, emotional stability, interpersonal and agreeableness, surgency and extraversion, business and openness to experience, leadership and surgency–extraversion).

Implications for Practice

In addition to scientific concerns related to defining the correct KSAOs for a leadership position are practice concerns related to gathering job-analytic information in a way that is efficient as well as accurate. Perhaps the most prevalent issue is whether a job analysis needs to be conducted at all. As with many other selection procedures, high-volume, high-stakes testing tends to drive careful job analysis. Even when the KSAO is known to be an effective predictor (e.g., cognitive ability), careful job analyses are conducted to enhance the defensibility of the procedure that will be used to evaluate a large number of candidates, often because of substantial mean group differences that result from the use of the test.

In contrast, lower volumes of candidates and jobs to be filled are less likely to be accompanied by extensive job analysis. Often the number of people who can speak to the requirements of a job is limited. For example, formal job analyses of a chief executive officer position for assessment purposes are rarely if ever done. Instead, an informal study may solicit opinions of what is needed for success from members of the board of directors and key leaders within the organization by an executive recruiter or human resources executive. Regardless of the difficulty of conducting the job analysis or its costs, a careful study of the actual requirements of the leadership role is necessary for both business

reasons and legal defensibility. Without some form of job analysis study, there is no way to know whether the right things are measured or whether the standards of acceptable performance are set at the appropriate levels for the organization.

Competency Models

Many organizations base their leadership assessments on their own competency models and do not conduct separate job analyses to support leadership assessment. Some of these models are grounded in careful analyses of work; others are based on corporate aspirations and expectations of employees. The more sophisticated competency models translate the competencies to specific behavioral expectations for various levels of employees. Although some are written in terms of individual traits or specific behavior expectations, many reflect a broad range of behaviors that in turn require a number of KSAOs for effective performance. For example, a common corporate competency is strategic focus, which consists of a set of behaviors ranging from envisioning the future to executing the strategy that require multiple KSAOs, including communication skills, business acumen, analytical skills, and so forth. When broad competencies are used to shape an assessment program, the practitioner must decide how best to measure them. Typically, the practitioner must choose between measures that attempt to evaluate the competency itself and measures that assess the component KSAOs. For example, strategic focus might be assessed through a business case that requires the candidate to develop a long-term business strategy and execution plan or through a 360-degree feedback instrument on which raters are asked to directly rate the candidate's level on the strategic focus competency. Alternatively, strategic focus might be decomposed into KSAOs that are necessary for such work, including cognitive ability, business and financial knowledge, and so forth. Often, both approaches are mingled. Because complex measures of company-defined competencies often do not exist, the challenge for the assessment professional is either to develop those broad-based simulations that mimic behaviors that are related to the competency, to create a 360-degree feedback instrument that evaluates competencies, or to identify a set of KSAOs that is related to that behavior and measure each KSAO directly. A major drawback to the KSAO approach can be the difficulty of explaining the relevance of the measures to the work of specific jobs, particularly for those instruments that are more abstract (e.g., a verbal reasoning test that measures cognitive ability). In addition, this approach does not effectively measure the individual's skill in applying the KSAOs he or she possesses to real-world problems.

Despite the desire for a common set of tasks and KSAOs for various leadership groups across organizations, the literature does not provide a great deal of support for such commonality. The published studies across many organizations are few, and the categorization of leadership positions is not rigorous. In addition, many experts in leadership (e.g., Fiedler, 1967; House, 1971) have emphasized the importance of context in shaping the requirements of the leadership role. Such a state of affairs emphasizes the importance of investigating the requirements of the leadership positions for which the assessment program is to be developed to ensure the right things are being measured.

LEADERSHIP ASSESSMENT TOOLS

A wide array of assessment tools can be used to evaluate many different leadership competencies and KSAOs. The next section provides an overview of the options and a brief discussion of the implications for practice.

Cognitive Ability Measures

Cognitive ability can be measured by many different tools ranging from objectively scored, multiple-choice tests to work samples that require a constructed response that is evaluated by trained assessors. Many of these work samples take realistic forms such as business cases, writing samples, or in-baskets and simulate actual work performed by leaders.

Research. The value of cognitive ability measures in predicting success in leadership positions is well established. In their meta-analytic work, Schmidt and Hunter (1998) reported an adjusted correlation of .58 between measures of cognitive ability and job

performance of managers. Further research (Hunter, Schmidt, & Judiesch, 1990) indicated that the complexity of the job defined in terms of information processing load moderates the relationship between cognitive ability and performance such that higher complexity is associated with larger correlations. Presumably, many managerial jobs fall into the category of jobs with higher information processing loads.

Implications for practice. Several concerns arise when using cognitive ability to evaluate the capabilities of leaders for selection or for development purposes; they are briefly discussed next.

From the point of view of an organization based in the United States, the likelihood of adverse impact for some protected classes is a significant concern when measuring cognitive ability. As discussed in Chapter 24 of this volume, meta-analytic studies of cognitive ability have consistently shown African Americans scoring about 1 standard deviation lower than Whites (Hough, Oswald, & Ployhart, 2001; Hunter & Hunter, 1984) and Hispanics scoring 0.5 to 0.8 standard deviations lower than Whites (Hough et al., 2001; Roth, Bevier, Bobko, Switzer, & Tyler, 2001), whereas Asians score about 0.2 standard deviations higher than Whites (Hough et al., 2001). In contrast, the differences between men and women tend to be small and inconsistent, although some evidence has shown that these differences depend on the type of cognitive ability measured as well as the method by which it was evaluated (Hyde, 2005; Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 1988). Most of these meta-analytic studies have not been limited to samples of leaders; however, the results are assumed to generalize to those kinds of positions. More important, some evidence has shown that increasing complexity decreases the *d* scores (Ones, Dilchert, Viswesvaran, & Salgado, 2010; Roth et al., 2001). This finding is attributed to self-selection and the minimum education requirements for many highly complex jobs. Again, many leadership roles are assumed to be complex jobs. Large mean group differences in scores on cognitive ability measures may limit an organization's ability to form a diverse leadership team and consequently make organizations reluctant to use such instruments. Nevertheless, it merits noting that the number of protected classes that are selected is dependent on the entire selection system, including the intercorrelations among predictors, the overall selection ratio, and the type of selection strategy used (e.g., multiple hurdles, top–down hiring) as well as the magnitude subgroup differences on the predictors (Sackett & Roth, 1996).

Individual participants' resistance to testing for cognitive abilities can be strong. Internal candidates for promotion and leaders being assessed for development may be quite concerned about sharing information about their intellectual abilities with people with whom they work, particularly if that capacity is limited. Often, measures of cognitive ability are somewhat abstract and appear academic. Both internal and external participants in leadership may question the relevance of such measures to leadership positions that do not involve certain tasks found in the job (e.g., numeric reasoning, algebra) and instead require application of mental abilities.

Many in organizations believe that evaluating cognitive abilities is not necessary because the range of cognitive ability in some populations, particularly executive populations, is highly restricted by virtue of the education these people are required to have and the experience they have acquired. However, research has not borne this out. Sackett and Ostgaard (1994) found a considerable amount of variability in cognitive ability even within groups with advanced degrees (e.g., doctoral, medical, and law degrees).

Occasionally, another challenge to using cognitive ability scores in leadership assessment is the belief that too much or too little cognitive ability is not useful. In other words, very bright people are not able to attend to the mundane and are quickly bored; less cognitively able participants simply cannot handle the challenges of leadership positions. A similar belief is the idea that the need for cognitive ability asymptotes at some point. This perspective holds that higher levels of cognitive ability are useful, but only to a point. If a curvilinear relationship between cognitive validity and performance in leadership positions did exist, significant challenges to establishing the validity of cognitive assessments and setting standards for leadership positions would also exist. Some (Bass, 1990) have argued

that cognitive ability may have a curvilinear relationship with performance, yet the research (Coward & Sackett, 1990) does not support this notion.

When an evaluation of cognitive abilities is included in a leadership assessment used for development purposes, the complaint may arise that an individual can do little about his or her cognitive abilities because they are firmly established by the time an adult is working and being considered for a leadership role. However, a documented lack of mental agility can pinpoint the nature of a deficit in problem solving and lead a participant to develop ways to compensate (e.g., relying on a talented team of subordinates).

A measure of high-level cognitive abilities can be expensive and challenging to develop and validate, and few instruments that evaluate very high-level abilities in a business context are commercially available. However, compared with other types of assessments (except personality inventories), most objective measures of cognitive ability are relatively inexpensive to administer.

Because questions on cognitive ability tests typically have a right and a wrong answer, many cognitive ability tests are administered in a proctored environment, even when they are used for development purposes. Although there are significant concerns about cheating on cognitive tests, many organizations have embraced unproctored Internet testing, and research on such testing continues (Tippins, 2009a, 2009c).

Personality Measures

Personality measures are frequently included in the assessment of leadership, and Ryan, McFarland, Baron, and Page (1999) reported an increase in the use of these measures. Although personality measures can take many forms, including objectively scored inventories and projective techniques, self-report measures in which the individual describes his or her own behavior are frequently used in leadership assessment.

Research. For jobs in general, meta-analytic research has indicated that personality variables do predict a variety of important criteria (Barrick, Mount, & Judge, 2001; Dudley, Orvis, Lebiecki, & Cortina, 2006; J. Hogan & Holland, 2003; J. Hogan & Ones, 1997; Hough & Ones, 2001; Hough & Oswald, 2008; Ones, Dilchert, Viswesvaran, & Judge, 2007; Ones, Viswesvaran, & Schmidt, 1993; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007; M. G. Rothstein & Goffin, 2006). Barrick et al. (2001) reported validities from meta-analytic research ranging from 0.07 to 0.27 for the Big Five personality constructs across broad occupational groups (Extraversion = 0.15; Conscientiousness = 0.27; Emotional Stability = 0.13; Agreeableness = 0.13; Openness to Experience = 0.07). Some have interpreted the overall validity estimates for the five factors in the Big Five model as demonstrating that personality tests have low validities with job performance (Morgeson et al., 2007). However, other researchers (Barrick et al., 2001; Ones et al., 2007) have noted that the validity for Conscientiousness (0.27) generalizes across all occupations studied, and the remaining four constructs predict "at least some criteria for some jobs" (Barrick et al., 2001, p. 22). For example, the validity of Extraversion predicting managerial performance is .21 (Barrick et al., 2001), and the validity for Agreeableness predicting performance in customer service jobs is .19 (Hurtz & Donovan, 2000).

Research results have also provided evidence for the validities of personality scales predicting managerial effectiveness. In addition to the Barrick et al. (2001) study mentioned in the preceding paragraph, Hough, Ones, and Viswesvaran (1998) found managerial effectiveness was predicted by Dominance (.27), energy level (.20), and achievement orientation (.17). Barrick and Mount (1991) found that Conscientiousness predicted managerial effectiveness (.22). In their summary of research on the validity of personality measures, Hough and Dilchert (2010) pointed out that higher validities are found when specific criteria are predicted by narrow personality traits that are theoretically aligned with the criteria. Oh and Berry (2009) demonstrated this finding by relating measures of the Big Five to ratings on a 360-degree feedback instrument that evaluated both task and contextual performance. The overall *R* of the five factors increased 74% (.23–.40) for task performance and 50% (.26–.39) for contextual performance. However, not all

research has shown acceptable validity coefficients for managerial jobs. Robertson, Barron, Gibbons, MacIver, and Nyfield (1993) found scores on a measure of conscientiousness did not predict managerial performance (.09) or promotability ratings (−.20).

Several factors in addition to the theoretical relevance of the criterion to the predictor likely affect the magnitude of the relationship between measures of personality and criteria of interest, including the type of criterion used, the method for measuring the criterion, the predictor method, the research setting, the research design (concurrent or predictive), item transparency, and rater perspective (Hough & Dilchert, 2010).

More important, most studies have indicated that the use of a personality inventory in an assessment program in combination with other measures increases the validity of the overall assessment. Bartram (2005) and McHenry, Hough, Toquam, Hanson, and Ashworth (1990) found incremental validity for personality measures when used with cognitive ability measures. DeGroot and Kluemper (2007) and McManus and Kelly (1999) found increments in validity over and above other measures, including interviews, biodata, and situational judgment tests.

Implications for practice. As with other assessment tools, personality inventories have strengths and weaknesses relative to the evaluation of leadership competencies. One concern is the possibility of nonlinear relationships between criteria and personality predictors—too much or too little of a personality trait is good or bad. However, the results are mixed. Day and Silverman (1989) and Robie and Ryan (1999) found no evidence of U-shaped relationships; however, Benson and Campbell (2007) found nonlinear relationships between "dark-side personality traits" as measured by the Global Personality Inventory and leadership performance measured in an assessment center and by the Hogan Development Survey and supervisory ratings of leadership performance.

In general, the addition of a personality measure to a leadership assessment program is not likely to limit the diversity of the selected group substantially. Most studies of racial and ethnic group differences and gender group differences have found trivial or inconsistent differences on most personality variables between Whites and protected racial subgroups and between men and women (Hough et al., 2001; Ones & Viswesvaran, 1998; see Chapter 38, this volume, for more information on group differences).

Despite the lack of large differences in scores across racial or gender groups, the use of personality tests introduces several problems in practical application. An important issue is that of faking. Because many personality inventories are self-report, the opportunity for faking is significant. Whether candidates for leadership positions actually do or do not fake, the opportunity to distort one's response brings into question the accuracy of an individual's score.

Another issue is that of the potential intrusiveness of some instruments. Although many personality inventories used in work settings use items that appear work related and are not particularly sensitive, anecdotally at least, some test takers believe some personality inventories pry and report feeling uncomfortable with the questions.

A special challenge for validating personality inventories for use in assessing leaders is the differences in results between concurrent and predictive studies that are often presumed to be related to the need to fake in a positive way. Validities are typically higher in concurrent studies than in predictive studies although there are exceptions. Because the faking problem is believed to be greater when the stakes are higher, the extent to which the results from a concurrent study can be generalized to a predictive study is not clear.

Many personality traits are believed to be formed early in life and are relatively but not completely stable. Roberts and DelVecchio (2000) indicated that personality trait consistency increases as an individual ages, peaking between ages 50 and 70. Nevertheless, when a leadership assessment is conducted for development purposes, the feasibility of changing one's personality sometimes arises. Because the exact mechanisms for changes in personality over time are not well understood, the question to be answered in developmental contexts is not so much "How do I change my personality?" as "How should I adapt my behavior given my natural tendencies?"

A personality inventory would be quite expensive to develop and research in light of the sample sizes required and the need to establish the meaning of scales through construct-oriented validity methods as well as the relationship to important criteria through criterion-oriented validation strategies; however, few practitioners do so unless for commercial purposes. A number of well-researched inventories that are appropriate for evaluating leaders and candidates for leadership positions are on the market.

As with cognitive ability instruments, the personality inventories are inexpensive to administer. As noted earlier, response distortions appear to weaken the relationship of personality scores to job performance criteria as typically evidenced in the higher validities found in concurrent studies compared with predictive studies, although this finding is not a consistent one. However, distorting responses intentionally or unintentionally does not appear to be affected by the presence or absence of an administrator. In fact, many personality inventories are delivered via the Internet in unproctored conditions.

Biodata

Biodata typically involves self-report descriptions of past experiences, behaviors, and attitudes. Most biodata inventories are objectively scored, and the scoring is based on empirical keys, rational keys, or a combination of both.

Research. Past experiences have been found to be highly predictive of performance across many types of jobs (Hunter & Hunter, 1984; Stokes, Mumford, & Owens, 1994) and for supervisors and managers in particular (Carlson, Scullen, Schmidt, Rothstein, & Erwin, 1999; Reilly & Chao, 1982; H. R. Rothstein, Schmidt, Erwin, Owen, & Sparks, 1990; Stokes & Cooper, 1994). H. R. Rothstein et al. (1990) reported a validity of .33 for biodata predicting managerial success. Dimensions of biodata that are often predictive of leadership success include academic achievement, family background, and economic stability and financial responsibility. In addition, personality-based biodata forms have been found to predict leadership potential (Stricker & Rock, 1998).

Implications for practice. Biodata forms can be particularly difficult to develop and validate. They share the same sample size problem as other forms of assessment in validity studies, but to some extent the sample size problem is even greater because of the number of people required to cross-validate scoring keys. However, a few biodata forms are commercially available, and Carlson et al. (1999) have demonstrated the generalizability of the scoring key across organizations and noted that validity did not vary greatly ($\rho = .53$, $SD = .05$).

Biodata may or may not limit the diversity of leaders in terms of race and gender depending on the instrument used. Meta-analyses have typically shown inconsistent results for differences in terms of the extent of the difference as well as the direction of the difference (see Chapter 25, this volume).

Several concerns about applicant reactions to biodata have emerged. The first is related to the candidates' ability to control certain aspects of their past life. For example, most children have little input into their parents' occupations. Similarly, some college students may have had little choice about working while pursuing an education and had little time available for extracurricular activities. The second is related to candidates' perception of the relevance of the measures to the job for which they are applying. For example, academic achievement in the past may not be obviously related to the ability to perform well in a sales management position in the future. Third, in a development context, the implication for change may not be clear. A low score on a biodata form may not clearly indicate what behaviors should change. Fourth, biodata items that are based on perceptions of one's own skills or others' perceptions of one's skills are subject to faking just as personality items are. Responses to transparent items can easily be faked and are likely to be distorted when the stakes are high as in a selection assessment. Requirements to elaborate on responses as well as the verifiability of the biodata items may deter faking (Ramsay, Schmitt, Oswald, Kim, & Gillespie, 2006; Schmitt et al., 2003).

There are firms that sell biodata forms at reasonable costs and have research bases to support their scoring. The cost of developing and validating a biodata form locally can be considerable if a scoring key

is part of the effort. As with other tests that can be administered via paper and pencil or computer, the costs of administration are relatively low. Proctors are not usually required because the kinds of distortions that are likely can occur in the presence of a proctor. Scoring of biodata forms almost always requires computer scoring procedures because of the complexity of the scoring algorithms.

Situational Judgment Tests

SJTs present a situation and several response options. The test-taker is asked to identify the best and worst response to the situation or to indicate the effectiveness of the response or the likelihood that he or she would take that action. Scoring keys can be developed in a number of ways, such as identifying what responses high performers typically make, determining which responses best separate high and low performers, and using subject matter experts to identify the effectiveness of response options.

Research. McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) estimated the validity of SJTs predicting job performance to be 0.34. They also noted a moderate relationship between SJTs and measures of cognitive ability ($\rho = .46$) indicating that other factors account for some of the variance in job performance. Bergman, Drasgow, Donovan, Henning, and Juraska (2006) found that the use of three different SJT scoring procedures (empirical = .25, subject matter expert = .32, hybrid participation = .17) produced test scores that predicted a six-item scale, Empowering Leadership, among a group of supervisors. Results from these scoring procedures and a fourth scoring procedure, a hybrid initiating structure key, provided significant incremental validity over cognitive ability and personality measures.

Implications for practice. SJTs typically reflect situations that the test taker is likely to face on the job and to some extent provide a realistic job preview. Consequently, there is little debate regarding the relevancy of the instrument. Moreover, "wrong" or less effective answers have obvious implications for development.

Research has suggested that SJTs that are used for selection may have some effect on the racial makeup of the leadership team. However, on the basis of the literature, SJTs would not be expected to have a significant impact on the number of women considered (see Chapter 30, this volume).

SJTs require significant effort to develop instruments appropriate for the positions for which they are used (Weekley & Ployhart, 2006). Few SJTs are commercially available, particularly for higher level leadership positions, and their generalizability across jobs and organizations has not been established. Relevant situations must be identified and realistic alternative actions developed and evaluated for effectiveness in the context in which the SJT will be used. Empirical scoring methods are particularly time and resource intensive. Once developed, however, administration is not particularly difficult. Whether SJTs should be proctored is still being researched and debated by practitioners.

One advantage to SJTs in practice is the type of validity studies that can be used. Criterion-related validation studies of SJTs are often difficult to execute, particularly when they are conducted with a managerial or executive sample, for the same reasons that cognitive ability tests are difficult to validate. The number of incumbents in the population is often too small for the statistics required; potential subjects are frequently unwilling to participate; and reliable evaluations of job performance are difficult to acquire. Evidence of validity for more abstract instruments such as measures of cognitive ability and personality and biodata forms may be based on criterion-related strategies because the relationship between the tests and the requirements of the job are less obvious. In contrast, SJTs often reflect the situations incumbents face, and evidence of validity may be established through a content strategy. Because content validation studies are usually much easier to execute, instruments whose validity evidence is based on a content-oriented strategy may be more desirable.

Structured Interviews

Structured interviews typically are composed of a set of questions developed to measure specific skills and often are accompanied by a set of probes to follow up or clarify a candidate's responses. The same set of questions is used for every candidate for a position

or in a specific assessment and development program. Another salient characteristic of structured interviews is the use of behavioral anchors that define the rating scale to evaluate responses. Typically, interviewers who use structured interviews are trained in appropriate interviewing techniques. In contrast, less structured interviews do not have a set of defined questions and tend to vary based on the interactions between the interviewer and the interviewee. Similarly, set standards against which to evaluate responses are usually not present. Instead, the interviewer makes a judgment based on internal standards using the information available.

Research. Interviews are a common part of most selection programs (McDaniel, Whetzel, Schmidt, & Maurer, 1994; Salgado, Viswesvaran, & Ones, 2001). However, structured interviews are not always part of a leadership assessment used for development. Often, a psychologist's interview, which is less structured, is used. Although the psychologist may begin with an interview protocol, he or she may deviate substantially to follow up on questions and probe more deeply into some areas. Frequently, no documented standards are available for evaluating the participant's responses, as in a structured interview, and the results are integrated with all the other information regarding the individual. This type of interview often solicits examples of behavioral tendencies that also might be detected in other instruments.

Little information is available on the validity of psychologists' interviews; indeed, many interviews result in no final predictions about the likelihood of future success. Instead, the results are used to describe personal characteristics and typical behaviors. In contrast, structured interviews have been studied extensively. McDaniel et al.'s (2001) meta-analysis of the employment interview indicated that the validity is .37.

Implications for practice. Interviews are an expected component of many leadership assessments. Anecdotally at least, many participants in both selection and development programs have reported expectations around describing their past experiences as well as their strengths and weaknesses. Consequently, the candidate population shows little resistance to interviews regardless of their form.

Research has suggested that interviews are unlikely to have substantial affects on the racial and gender composition of the leadership team when they are used for selection. Race and gender differences appear minimal or nonexistent (Harris, 1989; Huffcutt & Roth, 1998; see also Chapter 27, this volume).

The greatest disadvantages of the interview may be the need to set standards for interpretation and the skill required for effective interviewing. Interview questions are relatively easy to develop and validate, particularly when using a content-oriented approach; however, creating behavioral anchors can be difficult. Leadership roles may encompass a broad array of positions that range across many functions. Despite the differences in function, common skills or competencies may be necessary for effective leadership performance; however, the way in which these competencies are exhibited may vary considerably. Thus, very general anchors may be necessary if the competencies are to be used for all jobs. In addition, many practitioners struggle to get subject matter experts to differentiate acceptable performance from outstanding performance. Training interviewers to follow interview protocols and use the behavioral anchors provided can often also be challenging. The psychologist interview that is a part of many assessment programs requires considerably more skill than following an interview protocol and using behavioral anchors.

Relative to tests and inventories that can be administered via paper-and-pencil forms or computer software, interviews require the time of at least one interviewer and sometimes that of a panel. Often, the interviewer is a relatively high-level employee, for example, the manager of the position, peers of the position, or a professional psychologist. If the cost of training and calibration is factored into the equation, costs can go up precipitously. Consequently, interviews can be relatively expensive to administer because of the cost of the interviewer.

Individual Assessment

Individual assessment usually measures a variety of KSAOs, including cognitive ability and personality

traits, through several methods such as tests, inventories, 360-degree feedback, and interviews. The results are analyzed by an assessor who integrates the data and his or her understanding of the job requirements, provides an evaluation of the participant's relevant skills, and sometimes makes a decision about a candidate's fit to a position.

Research. The validity of the individual assessment is largely contingent on the validity of the individual instruments that make up the assessment, although the assessor's skill in integrating information and identifying trends may be an important contributor to the validity of overall results or recommendations. A combination of valid instruments is assumed to result in a valid assessment program. However, as noted earlier, little validity evidence is available for many predictors for higher level leadership roles because validity studies for these predictors are rarely done.

In addition to the problems of validating scores from single instruments, the rules for integrating data across instruments may not be specified. Often, such rules are quite complex and are contingent on the level of the score or the patterns of information. Thus, overall evaluations or decisions may not be based on the same information that is evaluated in the same manner. Moreover, the debate about the merits of clinical and statistical combination of the data continues (Kuncel & Highhouse, 2011; Silzer & Jeanneret, 2011). Most of the existing research (e.g., Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hazucha et al., 2011) has concluded that individual assessments are most predictive when data are combined mechanically instead of judgmentally. When data are combined by the assessor, the assessor's skill undoubtedly affects the accuracy of the final decision because the combination of test results is dependent on the assessor, who typically does not rely on formulaic rules for integrating data. Nevertheless, Korman (1968) and Prien, Schippmann, and Prien (2003) concluded that individual assessment has at least moderate validity. Recent meta-analytic work has investigated the relationship between individual psychological assessment and job performance; Roller and Morris (as reported in Silzer & Jeanneret, 2011) reported validity of .26. In contrast, Highhouse (2002) found little support for the validity of individual assessment:

> Very little research has been conducted on the efficacy of individual assessment practices, and the research base that does exist is vaguely described and outdated. As such, it is puzzling why individual assessment has not been subjected to scientific investigation in the way that other selection practices have. (p. 391)

Implications for practice. Individual assessments take on the advantages and disadvantages of their component parts. To the extent that highly educated candidates for high-level jobs feel insulted by a cognitive ability test or believe their privacy has been invaded by a personality inventory, the individual assessment process that incorporates those instruments may engender the same reactions. Similarly, instruments with large group mean differences will continue to reflect those differences when they are part of an individual assessment program.

An interesting finding is that candidates seem to have more confidence in individual assessments when they are administered by an external psychologist than when the same or similar instrument is administered by the human resources representatives in their organization. Perhaps participants are placing great faith in the external psychologist's ability to combine disparate data into a whole picture of the individual. Although the use of an external consultant can build confidence in the assessment, it will also add to its cost and occasionally limit the flexibility of scheduling of the assessment.

Individual assessments enjoy a great deal of popularity, especially in situations that involve a high-level position and for which the stakes are high for both the participant and the organization. Some of that popularity likely stems from the need of both parties, the participant and the organization, to avoid making a catastrophic mistake. Another contributor to the acceptability of individual assessment is the minimal upfront work that is required before the assessment. Often, common leadership competency requirements are assumed, and a brief conversation about the organization's culture, the role, and its special challenges is sufficient to give the assessor

an idea of the characteristics of a candidate who will be successful in the job.

Costs to develop an individual assessment can be substantial if a diligent approach to establishing the job requirements and a careful review of potential instruments are undertaken. Because these kinds of analyses are often reusable, costs for the development of an individual assessment program are often amortized over multiple uses. Administration costs vary considerably depending on the types of instruments used and the type of administrator required.

360-Degree Feedback

Multirater or 360-degree feedback instruments solicit information about an individual's skills or performance from a variety of sources, including his or her manager, direct reports, peers, and other stakeholders and are sometimes used to evaluate leaders. A 360-degree feedback instrument may be used alone or in combination with other methods for assessing individual capability or with performance data. A great deal of controversy exists regarding the appropriateness of using 360-degree feedback instruments for selection purposes. Concerns about the accuracy of the ratings of the target individual and those of the other raters as well as the legal defensibility of the ratings call into question the wisdom of their use.

Research. Another concern is the validity of making inferences about future behavior on the basis of past performance in dissimilar positions. For example, accurate ratings of a person in a managerial role that requires extensive interactions with direct reports may not be predictive of that individual's ability to develop a strategy for building talent in the organization in a higher level, executive position. In other words, 360-degree feedback instruments tend to focus on the individual's past behavior. Although past behavior often predicts future behavior, the past behavior must be related to the future behavior if the prediction is to be accurate, and a 360-degree feedback instrument that is focused on the requirements of one position level may not provide sufficient information about what the person is capable of in the future in a higher level position.

The published research regarding the criterion-related validity of 360-degree ratings has provided some evidence of their predictiveness. Halverson, Tonidandel, Barlow, and Dipboye (2005) found that ratings and self–other agreement predicted promotion rate in the U.S. Air Force. Darr and Catano (2008) found that supervisor (.34) and peer (.23) ratings of senior managers predicted their performance in a selection interview. Certainly, manager ratings can be expected to be predictive if for no other reason than most managers' substantial roles in the promotion of their subordinates.

Implications for practice. Perhaps the greatest challenge to using multirater instruments is the administrative hassle of identifying raters and ensuring their thoughtful attention. When 360-degree evaluations are deployed widely within one organization, individuals at the top of the hierarchy can find themselves with multiple questionnaires to complete.

In addition to getting the raters' attention long enough to complete the form, the practitioner must also consider how best to obtain accurate ratings. As with all self-report instruments, the focal individual may intentionally or unintentionally distort his or her ratings. Despite promises of confidentiality and anonymity, other raters may worry that negative ratings will have a direct or indirect effect on them and adjust their ratings accordingly. In extreme situations, some may even want to encourage particular decisions through positive ratings (e.g., promote an incompetent manager) or negative ratings (e.g., retain a good manager).

Assessment Centers

Assessment centers are designed almost exclusively for the evaluation of leadership talent. As with individual assessment programs, they usually combine a number of techniques, although the defining characteristics of the assessment center are the direct observation of performance in structured situations, ratings on multiple competencies established through job analysis by multiple observers, and systematic procedures for recording, integrating, and summarizing candidates' behaviors.

Research. Previous research has converged on similar values for criterion-related evidence of validity. In their review of meta-analytic studies, Schmidt and

Hunter (1998) reported the validity of the assessment center to be .37 on the basis of the work of Gaugler, Rosenthal, Thornton, and Bentson (1987). Arthur, Day, McNelly, and Edens (2003) reported a validity of .36. Thornton and Rupp (2006) reported validity estimates for assessment center ratings and management success ranging from .31 to .42. However, all of the meta-analytic estimates have considerable variability, suggesting that assessment centers range considerably in predictive power.

One long-standing controversy regarding assessment centers and their validity is the construct validity problem. Although most assessment centers use multiple measures to evaluate a behavioral dimension, convergent validity via multitrait–multimethod model is difficult to attain, and exercises appear to be the better predictors. Lance (2008) has argued that the behavior of participants in assessment centers is cross-situationally specific and the multitrait–multimethod model is inappropriate because the exercise effects that are commonly found are the result of true cross-exercise differences. Thus, exercise ratings rather than assessment center ratings or dimensions should be the predictor of interest. Yet, others (Connelly, Ones, Ramesh, & Goff, 2008; Howard, 2008; Rupp, Thornton, & Gibbons, 2008) have noted some cross-situational consistency and argued in favor of assessment center dimensions. Connelly et al. (2008) and Melchers and Konig (2008) have described the meta-analytic evidence supporting dimensions in assessment centers. In addition, Rupp et al. (2008) have pointed out that all human performance, including performance in assessment centers, is multidimensional. Undoubtedly, the ongoing controversy will stimulate further research on topics including effective ways to define dimensions, measure behavior in the assessment center, train assessors, and so forth.

Arthur et al. (2003) demonstrated that the various dimensions used in assessment centers could be collapsed into seven dimensions (consideration–awareness of others, communication, drive, influencing others, organizing and planning, problem solving, stress tolerance) and provided evidence of the criterion-related validity of those dimensions for predicting job performance. Bowler and Woehr (2006) developed a research-based model of assessment center dimensions based on both dimension and exercise factors and demonstrated that exercise effects were, on average, 34% of variance in a single assessment center rating, whereas dimension variance constituted 22% of variance.

Assessment centers use multiple measures to provide an assessment of an individual's competencies. Consequently, assessment center results are likely to reflect the typical group differences of the individual procedures that compose it. Although group difference information is evaluated for various measures in other chapters, the group difference literature for assessment centers is reviewed here.

In their meta-analytic work, Dean, Roth, and Bobko (2008) found Black–White differences of .52 and Hispanic–White differences of .28. Other researchers have noted that subgroup differences based on race in assessment center ratings appear to be associated with cognitively loaded exercises (Hough et al., 2001). When ratings are based on exercises that are cognitively loaded, group mean differences based on race are commonly found (Goldstein, Yusko, Braverman, Smith, & Chung, 1998). African Americans score lower than Whites. For example, African Americans and Whites scored approximately the same on a subordinate meeting assessment, whereas Whites scored higher than African Americans on other assessment exercises. The group discussion and project presentation exercises had about .25 standard deviation difference. The African American–White d values varied from -0.35 to -0.40 for the in-basket, in-basket coaching, project discussion, and team presentation exercises. The African American–White d value for the overall assessment center score was -0.40. Bobrow and Leonards (1997) noted that when ratings are based on exercises that are more focused on interpersonal skills, no differences between Whites and minorities were typically found.

Comparisons of men's and women's performance in assessment centers has produced inconsistent results in the literature. Several studies found no significant differences in promotion ratios and middle-management potential (Alexander, Buck, & McCarthy, 1975; Moses, 1973; Ritchie & Moses, 1983). L. R. Anderson and Thacker (1985) found no differences in the overall assessment rating, but

Walsh, Weinberg, and Fairfield (1987); Bobrow and Leonards (1997), and Dean et al. (2008) found women received higher ratings on the overall assessment rating. At more specific rating levels, a varying pattern has also emerged. Shore (1992) found differences favoring women on performance-style skills; Schmitt (1993) found performance differences favoring women on all dimensions; Shore, Tashchian, and Adams (1997) concluded that there were no significant differences between the men's and women's scores on role-play exercises. More recently, N. Anderson, Lievens, van Dam, and Born (2006) found that women were rated higher on constructs related to interpersonally oriented leadership styles (i.e., oral communication and interaction) and drive and determination.

Implications for practice. As with other approaches to assessing leaders, assessment centers have both advantages and disadvantages. From the perspective of many, the primary advantage of this approach is its realism. In well-developed assessment centers, the tasks, exercises, and role plays assigned to the assessee are obviously related to the leadership role for which he or she is being assessed. This realism tends to affect candidate reactions positively and lead participants to accept readily the developmental suggestions based on the assessment results. At the same time, that realism is the source of the major disadvantage, the cost of developing and administering the assessment center. The development of appropriate simulations is time consuming, and administration can require role players as well as multiple assessors and adequate physical facilities.

The costs relative to the benefits is an important issue practitioners must face when considering assessment centers. Given the high costs of assessment centers, many practitioners have questioned their value. However, several researchers have found incremental validity in predicting job performance beyond measures of cognitive ability or personality. Goldstein et al. (1998) found incremental validity over and above measures of cognitive ability. Using meta-analysis, Meriac, Hoffman, Woehr, and Fleisher (2008) demonstrated that Arthur et al.'s (2003) seven dimensions accounted for variance in job performance beyond that accounted for by cognitive ability and personality variables. Looking at executive managers (candidates for manager of a German police department), Krause, Kersting, Heggestad, and Thornton (2006) also found incremental validity over measures of cognitive ability alone. Other intangible benefits have also been noted relative to the participant's perceptions of the validity of the results and the likelihood that he or she will act on those results in developmental assessment centers.

When using the assessment center method for leadership assessment, the practitioner must carefully consider the assessment center characteristics. Thornton and Krause (2009) found distinct differences between assessment centers used for selection and development. For example, assessment centers used for selection tended to emphasize intellectual and problem-solving skills, and those used for development focused more on interpersonal and intrapersonal skills.

The extent to which the use of assessment center results for selection purposes will limit the diversity of the candidate pool appears to depend heavily on the type of exercises included and the dimensions that are measured.

Measures of Fit

Fit can be defined in terms of congruence between the leader and others, and the degree to which an individual fits with his or her organization, team, or job is often an important element in the selection of leaders, particularly at the higher levels. Fit can also be defined in terms of a match between the needs of the organization and the capabilities of the leader. For example, when the situation calls for a focus on accomplishing work, a leader who is task oriented would be hypothesized to be more successful than one who is relationship oriented.

Research. Research has indicated that fit measures have some validity. Posner, Kouzes, and Schmidt (1985) found that congruence on values between managers and their organization was positively related to the managers' success and intention to remain in the organization. Weiss (1978) found that congruence on values between supervisors and their subordinates predicted the subordinates' ratings of

the supervisors' competence and success. Fleenor, McCauley, and Brutus (1996) investigated the relationship between agreement in self- and subordinate ratings and leader effectiveness. When level of performance was controlled for, the groups showed no differences between them defined by level of agreement.

Implications for practice. One of the major drawbacks to evaluating an individual for fit with a team, a culture, or a particular challenge is the likelihood that things will change. In a rapidly changing world, few organizations face the same demands year after year. Thus, one of the drawbacks to selecting a leader who matches the situation at a particular point in time is that the situation may well change. The work of Pulakos, Arad, Donovan, and Plamandon (2000) has suggested that those who are most adaptable are most likely to succeed in leadership roles. Ensuring the fit of a leader may also be counterproductive if the leader is not able to adapt when radical change is required. Some evidence has also suggested that close fit may limit creativity and consideration of alternative approaches to difficult problems and that complementary skills may be result in better solutions to problems. A related question is whether the existing management team or culture should be the basis for fit or new ideas and fresh approaches would be useful.

Another problem with using measures of fit is the difficulty of defining what characteristics of entities such as an organization or a team or even the culture are relevant and then determining what characteristics of the individual leader are needed to meet those demands successfully. Extensive research may be necessary to determine the relevant variables to be measured, and the practitioner risks obtaining the results just as the requirements change.

SUMMARY

When developing leadership assessments, practitioners have a variety of options from which to choose, each with its advantages and disadvantages relative to the purpose for which it will be used. Regardless of the tools chosen to evaluate the capabilities of leaders and candidates for leadership roles, effective leadership assessment requires a multistep process. First, the psychologist must understand the purposes of the assessment and the likely population of people with whom it will be used. Second, the competency and KSAO requirements of the leadership positions for which the assessment will be used must be fully understood. Third, assessment tools that are appropriately aligned with the constructs to be measured and the purpose of the assessment must be identified or developed. In addition, the tools must take into consideration the population with which they will be used as well as the setting in which the assessment will be administered. Fourth, the psychologist should ensure there is a rationale for the use and interpretations of the assessment procedures. When an assessment program cannot be validated using typical strategies, alternative validation procedures should be considered (see McPhail, 2007). Fifth, implementation should be carefully considered to ensure fair treatment of participants that is consistent with the purpose of the assessment. Finally, the psychologist responsible for the assessment program must monitor and evaluate it to ensure its usefulness for the stated purposes.

References

Alexander, H. S., Buck, J. A., & McCarthy, R. J. (1975). Usefulness of the assessment center process for selection to upward mobility programs. *Human Resource Management, 75*, 11–13.

Anderson, L. R., & Thacker, J. (1985). Self-monitoring and sex as related to assessment center ratings and job performance. *Basic and Applied Social Psychology, 6*, 345–361. doi:10.1207/s15324834basp0604_5

Anderson, N., Lievens, F., van Dam, K., & Born, M. (2006). A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal of Applied Psychology, 91*, 555–566. doi:10.1037/0021-9010.91.3.555

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125–153. doi:10.1111/j.1744-6570.2003.tb00146.x

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where

do we go next? *International Journal of Selection and Assessment, 9*, 9–30. doi:10.1111/1468-2389.00160

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203. doi:10.1037/0021-9010.90.6.1185

Bass, B. M. (1990). From transactional to transformational leadership: Learning to share the vision. *Organizational Dynamics, 18*, 19–31. doi:10.1016/0090-2616(90)90061-S

Benson, M. J., & Campbell, J. P. (2007). To be, or not to be, linear: An expanded representation of personality and its relationship to leadership performance. *International Journal of Selection and Assessment, 15*, 232–249. doi:10.1111/j.1468-2389.2007.00384.x

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*, 223–235. doi:10.1111/j.1468-2389.2006.00345.x

Bobrow, W., & Leonards, J. S. (1997). Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality, 12*, 217–236.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*, 1114–1124. doi:10.1037/0021-9010.91.5.1114

Carlson, K. D., Scullen, S. E., Schmidt, F. L., Rothstein, H., & Erwin, F. (1999). Generalizable biographical data validity can be achieved without multi-organizational development and keying. *Personnel Psychology, 52*, 731–755. doi:10.1111/j.1744-6570.1999.tb00179.x

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 121–124. doi:10.1111/j.1754-9434.2007.00022.x

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability–performance relationships: A reconfirmation. *Journal of Applied Psychology, 75*, 297–300. doi:10.1037/0021-9010.75.3.297

Darr, W., & Catano, V. M. (2008). Multisource assessments of behavioral competencies and selection interview performance. *International Journal of Selection and Assessment, 16*, 68–72. doi:10.1111/j.1468-2389.2008.00410.x

Davies, S., Hogan, J., Foster, J., & Elizondo, F. (2005, April). *Recombinant personality measures for predicting leadership competence.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42*, 25–36. doi:10.1111/j.1744-6570.1989.tb01549.x

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93*, 685–691. doi:10.1037/0021-9010.93.3.685

DeGroot, T., & Kluemper, D. (2007). Evidence of predictive and incremental validity of personality factors, vocal attractiveness and the situational interview. *International Journal of Selection and Assessment, 15*, 30–39. doi:10.1111/j.1468-2389.2007.00365.x

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57. doi:10.1037/0021-9010.91.1.40

Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*, 38290–38315.

Erker, S. C., Cosentino, C. J., & Tamanini, K. B. (2010). Selection methods and desired outcomes: Integrating assessment content and technology to improve entry- and mid-level leadership performance. In J. F. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 721–740). New York, NY: Routledge.

Fiedler, F. (1967). *A theory of leadership effectiveness.* New York, NY: McGraw-Hill.

Fleenor, J. W., McCauley, C. D., & Brutus, S. (1996). Self–other agreement and leader effectiveness. *Leadership Quarterly, 7*, 487–506. doi:10.1016/S1048-9843(96)90003-X

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493–511. doi:10.1037/0021-9010.72.3.493

Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51*, 357–374. doi:10.1111/j.1744-6570.1998.tb00729.x

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction. *Psychological Assessment, 12*, 19–30. doi:10.1037/1040-3590.12.1.19

Halverson, S. K., Tonidandel, S., Barlow, C., & Dipboye, R. L. (2005). Self–other agreement on a 360-degree leadership evaluation. In S. Reddy (Ed.), *Perspectives*

on multirater performance assessment (pp. 125–144). Hyderabad, India: ICFAI.

Harris, M. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology, 42*, 691–726. doi:10.1111/j.1744-6570.1989.tb00673.x

Hazucha, J. F., Ramesh, A., Goff, M., Crandell, S., Gerstner, C., Sloan, E., . . . Van Katwyk, P. (2011). Individual psychological assessment: The poster child of blended science and practice. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 297–301. doi:10.1111/j.1754-9434.2011.01342.x

Hemphill, J. K. (1960). *Dimensions of executive positions.* Oxford: Ohio State University Press.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55*, 363–396. doi:10.1111/j.1744-6570.2002.tb00114.x

Hogan, J., Davies, S., & Hogan, R. (2007). Generalizing personality-based validity evidence. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 181–229). Hoboken, NJ: Wiley.

Hogan, J., & Holland, H. (2003). Using theory to evaluate personality and job–performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100–112. doi:10.1037/0021-9010.88.1.100

Hogan, J., & Ones, D. S. (1997). Conscientiousness and integrity at work. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 849–870). London, England: Academic Press.

Hogan, R., & Warrenfeltz, R. (2003). Educating the modern manager. *Academy of Management Learning and Education, 2*, 74–84. doi:10.5465/AMLE.2003.9324043

Hough, L. M., & Dilchert, S. (2010). Personality: Its measurement and validity for employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 299–319). Mahwah, NJ: Erlbaum.

Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology: Vol. 1. Personnel psychology* (pp. 233–277). Thousand Oaks, CA: Sage. doi:10.4135/9781848608320.n13

Hough, L. M., Ones, D. S., & Viswesvaran, C. (1998). Personality correlates of managerial performance constructs. In R. C. Page (Chair), *Personality determinants of managerial potential performance, progression and ascendancy*. Symposium conducted at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial–organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 272–290. doi:10.1111/j.1754-9434.2008.00048.x

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194. doi:10.1111/1468-2389.00171

House, R. J. (1971). A path-goal theory of leader effectiveness. *Administrative Science Quarterly, 16*, 321–339. doi:10.2307/2391905

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 98–104. doi:10.1111/j.1754-9434.2007.00018.x

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83*, 179–189. doi:10.1037/0021-9010.83.2.179

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*, 28–42. doi:10.1037/0021-9010.75.1.28

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–879. doi:10.1037/0021-9010.85.6.869

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581–592. doi:10.1037/0003-066X.60.6.581

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139–155. doi:10.1037/0033-2909.107.2.139

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*, 53–69. doi:10.1037/0033-2909.104.1.53

Korman, A. K. (1968). The prediction of managerial performance: A review. *Personnel Psychology, 21*, 295–322. doi:10.1111/j.1744-6570.1968.tb02032.x

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*, 360–371. doi:10.1111/j.1468-2389.2006.00357.x

Kuncel, N. R., & Highhouse, S. (2011). Complex predictions and assessor mystique. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 302–306. doi:10.1111/j.1754-9434.2011.01343.x

Lance, C. E. (2008). Why assessment centers (AC) don't work the way they're supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 84–97. doi:10.1111/j.1754-9434.2007.00017.x

Mahoney, T. A., Jerdee, T. H., & Carroll, S. J. (1965). The job(s) of management. *Industrial Relations, 4*, 97–110. doi:10.1111/j.1468-232X.1965.tb00922.x

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740. doi:10.1037/0021-9010.86.4.730

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616. doi:10.1037/0021-9010.79.4.599

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335–354. doi:10.1111/j.1744-6570.1990.tb01562.x

McManus, M. A., & Kelly, M. L. (1999). Personality measures and biodata evidence regarding their incremental predictive value in the life insurance industry. *Personnel Psychology, 52*, 137–148. doi:10.1111/j.1744-6570.1999.tb01817.x

McPhail, S. M. (Ed.). (2007). *Alternative validation strategies: Developing new and leveraging existing validity evidence*. New York, NY: Pfeiffer.

Melchers, K. G., & Konig, C. J. (2008). It is not yet time to dismiss dimensions in assessment centers. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 125–127. doi:10.1111/j.1754-9434.2007.00023.x

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042–1052. doi:10.1037/0021-9010.93.5.1042

Mintzberg, H. (1975). The manager's job: Folklore and fact. *Harvard Business Review, 53*, 100–110.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729. doi:10.1111/j.1744-6570.2007.00089.x

Moses, J. L. (1973). The development of an assessment center for the early identification of supervisory potential. *Personnel Psychology, 26*, 569–580. doi:10.1111/j.1744-6570.1973.tb01158.x

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2007). The leadership skills strataplex: Leadership skill requirements across organizational levels. *Leadership Quarterly, 18*, 154–166. doi:10.1016/j.leaqua.2007.01.005

Oh, I.-S., & Berry, C. M. (2009). The five-factor model of personality and managerial performance: Validity gains through the use of 360 degree performance ratings. *Journal of Applied Psychology, 94*, 1498–1513. doi:10.1037/a0017221

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027. doi:10.1111/j.1744-6570.2007.00099.x

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2010). Cognitive ability. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 255–276). Mahwah, NJ: Erlbaum.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245–269.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679–703. doi:10.1037/0021-9010.78.4.679

Pearlman, K., & Sanchez, J. (2010). Work analysis. In J. F. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 73–98). New York, NY: Routledge.

Posner, B., Kouzes, J., & Schmidt, W. (1985). Shared values make a difference: An empirical test of corporate culture. *Human Resource Management, 24*, 293–309. doi:10.1002/hrm.3930240305

Prien, E. P., Schippmann, J. S., & Prien, K. O. (2003). *Individual assessment as practiced in industry and consulting*. Mahwah, NJ: Erlbaum.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*, 612–624. doi:10.1037/0021-9010.85.4.612

Ramsay, L. J., Schmitt, N., Oswald, F. L., Kim, B. H., & Gillespie, M. A. (2006). The impact of situational context variables on responses to biodata and situational judgment inventory items. *Psychology Science, 48*, 268–287.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures.

Personnel Psychology, 35, 1–62. doi:10.1111/j.1744-6570.1982.tb02184.x

Ritchie, R. J., & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology, 68*, 227–231. doi:10.1037/0021-9010.68.2.227

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*, 3–25. doi:10.1037/0033-2909.126.1.3

Roberts, B. W., Kuncel, N., Shiner, R. N., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socio-economic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*, 313–345. doi:10.1111/j.1745-6916.2007.00047.x

Robertson, I. T., Barron, H., Gibbons, P., MacIver, R., & Nyfield, G. (1993). Conscientiousness and managerial performance. *Journal of Occupational and Organizational Psychology, 66*, 225–244. doi:10.1111/j.2044-8325.1993.tb00534.x

Robie, C., & Ryan, A. M. (1999). Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. *International Journal of Selection and Assessment, 7*, 157–169. doi:10.1111/1468-2389.00115

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297–330. doi:10.1111/j.1744-6570.2001.tb00094.x

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owen, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75*, 175–184. doi:10.1037/0021-9010.75.2.175

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180. doi:10.1016/j.hrmr.2006.03.004

Rupp, D. E., Thornton, G. C., III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 116–120. doi:10.1111/j.1754-9434.2007.00021.x

Ryan, A. M., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52*, 359–392. doi:10.1111/j.1744-6570.1999.tb00165.x

Sackett, P. R., & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79*, 680–684. doi:10.1037/0021-9010.79.5.680

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49*, 549–572. doi:10.1111/j.1744-6570.1996.tb01584.x

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology: Vol. 1. Personnel psychology* (pp. 165–199). Thousand Oaks, CA: Sage. doi:10.4135/9781848608320.n10

Schippmann, J. S., Prien, E. P., & Hughes, G. L. (1991). The content of management work: Formation of task and job skill composite classifications. *Journal of Business and Psychology, 5*, 325–354. doi:10.1007/BF01017706

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmitt, N. (1993). Group composition, gender, and race effects on assessment center ratings. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 315–332). Hillsdale, NJ: Erlbaum.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88*, 979–988. doi:10.1037/0021-9010.88.6.979

Shore, T. H. (1992). Subtle gender bias in the assessment of managerial potential. *Sex Roles, 27*, 499–515. doi:10.1007/BF00290006

Shore, T. H., Taschian, A., & Adams, J. S. (1997). The role of gender in a developmental assessment center. *Journal of Social Behavior and Personality, 12*, 191–203.

Silzer, R. (2002). *The 21st century executive: Innovative practices for building leadership at the top.* San Francisco, CA: Jossey-Bass.

Silzer, R., & Jeanneret, P. R. (2011). Individual psychological assessment: A practice and science in search of common ground. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 270–296. doi:10.1111/j.1754-9434.2011.01341.x

Stokes, G. S., & Cooper, L. A. (1994). Selection using biodata: Old notions revisited. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook* (pp. 311–349). Palo Alto, CA: CPP Books.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (Eds.). (1994). *Biodata handbook*. Palo Alto, CA: CPP Books.

Stricker, L. J., & Rock, D. A. (1998). Assessing leadership potential with a biographical measure of personality traits. *International Journal of Selection and Assessment, 6*, 164–184. doi:10.1111/1468-2389.00087

Thornton, G. C., & Krause, D. E. (2009). Selection versus development assessment centers: An international survey of design, execution, and evaluation. *International Journal of Human Resource Management, 20*, 478–498. doi:10.1080/09585190802673536

Thornton, G. C., & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Erlbaum.

Tippins, N. T. (2009a). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 2–10. doi:10.1111/j.1754-9434.2008.01097.x

Tippins, N. T. (2009b, April). *Selection of first line supervisors: What we know*. Workshop presented at the meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Tippins, N. T. (2009c). Where is the unproctored Internet testing train headed now? *Industrial and Organizational Psychology: Perspectives on Science and Practice,* 2, 69–76. doi:10.1111/j.1754-9434.2008.01111.x

Tornow, W. W., & Pinto, P. R. (1976). The development of a managerial job taxonomy: A system for describing, classifying, and evaluating executive positions. *Journal of Applied Psychology, 61*, 410–418. doi:10.1037/0021-9010.61.4.410

Walsh, J. P., Weinberg, R. M., & Fairfield, M. L. (1987). The effects of gender on assessment centre evaluations. *Journal of Occupational Psychology, 60*, 305–309. doi:10.1111/j.2044-8325.1987.tb00262.x

Weekley, J. A., & Ployhart, R. E. (2006). *Situational judgment tests: Theory, measurement, and application.* Mahwah, NJ: Erlbaum.

Weiss, H. M. (1978). Social learning of work values in organizations. *Journal of Applied Psychology, 63*, 711–718. doi:10.1037/0021-9010.63.6.711

CHAPTER 27

UNDERSTANDING AND IMPROVING EMPLOYEE SELECTION INTERVIEWS

Robert L. Dipboye and Stefanie K. Johnson

Among the most important applications of psychological assessment is the selection of employees for jobs. Most of the attention in the employee selection research has been on the tests and inventories designed to measure the knowledge, skills, abilities, and other characteristics (KSAOs) required for a given job. This research has led to sophisticated assessment tools with the potential to contribute substantially to selection decisions (Schmidt & Hunter, 1998). What has been largely ignored, however, is the social process in which these instruments are embedded. Not only is the face-to-face interview the most popular approach to assessing candidates for employment (e.g., Furnham, 2008; Harris, Dworkin, & Park, 1990), but it is also at the core of the social and decision-making processes through which employees are assessed, screened, and selected in many organizations. For example, it is not uncommon for the scores on other selection instruments, such as mental ability tests (see Chapter 24, this volume), personality inventories (see Chapter 28, this volume), and biographical data (see Chapter 25, this volume), to be made available to interviewers, who then use them as basis for their questions and incorporate them into their final evaluations. In these situations, tests and other objective assessment tools influence the lens through which interviewers perceive candidates, ultimately influencing interviewer judgments.

In this chapter, we review the research on interviewing as used to assess applicants for employment. We start by examining the evaluation of how interviewer judgments perform as a tool of assessment in employee selection. Next, we examine the interview as a multiphase process that starts with the interviewer's and applicant's first encounter in the previewing of paper credentials and in the first few minutes of the interview, which we call the *preinterview phase*. Then, we discuss the details of what occurs during the interview itself, during which the interviewer gathers and processes information about the applicant, which we call the *interview phase*. The third and last interview phase involves the final judgment of the applicant's qualifications, which we call the *postinterview phase*. We review the research related to each of these three phases of the interview (preinterview, interview, postinterview) and the alternative processes that can be used to describe the linkages between prior impressions and the subsequent gathering and processing of information. We conclude with some suggestions for improving the psychometric quality of interviewer assessments.

EVALUATING STRUCTURED SELECTION INTERVIEWS

Over the past century, a considerable amount of research has been conducted to evaluate how well the interview performs as a tool for assessing job applicants. The meta-analyses of these research findings have focused on validity and reliability and have clearly shown that interviewer judgments can be of great value in the selection process.

DOI: 10.1037/14047-027
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

Criterion-Related Validity and Reliability of Structured Interviews

How well interviewer judgments perform in selecting employees depends on the structure of the procedures used by the interviewer. Differences have occurred in how researchers have defined interview structure (Campion, Palmer, & Campion, 1997; Chapman & Zweig, 2005; Conway, Jako, & Goodman, 1995; Dipboye, Wooten, & Halverson, 2004; Huffcutt & Arthur, 1994). However, the consensus seems to be that at a minimum a structured interview is characterized by the use of (a) the same questions with all applicants, (b) job-related questions, and (c) a scoring protocol and numerical rating scales for evaluating applicant responses to these questions. At the other extreme, interviewers using unstructured procedures can ask whatever questions they deem important, in whatever order they want to ask them. The purpose is usually to get a sense of who the applicant is as a total person. If rating scales are used in unstructured interviews, they tend to be focused on global evaluations of the applicant rather than on specific, job-related dimensions.

Meta-analyses of interviewer judgments of applicants have consistently shown that structured interviews yield validities and reliabilities that can rival the levels demonstrated for mental ability tests, work samples, and scored biographical inventories (Schmidt & Zimmerman, 2004), with highly structured interviews boasting average validities of .56 (Campion, Pursell, & Brown, 1988). The corrected criterion-related validities of structured interviews reported in meta-analyses have ranged from .44 (McDaniel, Whetzel, Schmidt, & Maurer, 1994) to .63 (Wiesner & Cronshaw, 1988).

Structured interviews also appear to be highly reliable. Compared with unstructured interviews, which have reliabilities in the .60s, meta-analyses have revealed that structured interviews have reliabilities in the .80s (McDaniel et al., 1994; Wiesner & Cronshaw, 1988). Conway et al. (1995) found that reliabilities were higher when panel interviews were used (.77) than when standard one-on-one interviews were used (.53). However, their moderator analyses showed that each of these estimates was stronger when the interview was structured.

Legal Advantages of Selection Interviews

In addition to achieving higher validities and reliabilities than unstructured interviews, structured interviews also appear to provide possible legal advantages. Equal Employment Opportunity Commission (1978) guidelines require that all selection devices, including interviews, do not unfairly discriminate against applicants on the basis of race, sex, color, religion, national origin, disability, pregnancy, or age older than 40. Race can never be used as a fundamental requirement of the job and, except for rare situations, neither can the other characteristics. The only exception is if they can be shown to be fundamental requirements of the job or bona fide occupational qualifications. Under the legal principle of disparate impact, if a selection device resulted in a statistical disparity (consistent with the four-fifths rule) against applicants from a group protected under the law, the organization would need to demonstrate that the selection device is valid or of business necessity. The research evaluating adverse impact in the interview is mixed with some research showing small or no mean differences between Whites and racial minorities (Sackett & Ellingson, 1997) and other research showing substantial mean differences (Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002). The studies examining bias in prediction for structured interviews are too few to provide a strong basis for a conclusion (cf. Campion et al., 1988; Pulakos & Schmitt, 1995). The research that exists has shown that structured interviews are equally valid across subgroups. For instance, Pulakos and Schmitt (1995) found that an experience-based structured interview was equally valid for White, Black, Hispanic, male, and female subgroups. Campion et al. (1988) found some intercept differences for race on a structured interview but no differences in slope for either race or sex. Even if there is adverse impact, the interview structure was consistent with the hiring guidelines set down in case law and in the Equal Employment Opportunity Commission guidelines. Consequently, structured interviews can provide the basis for a stronger legal defense in the event of a discrimination suit than can unstructured interviews (Williamson, Campion, Malos, Roehling, & Campion, 1997).

Unresolved Issues

Despite the evidence favoring structured over less structured interviews, several crucial issues remain unresolved. One of these is whether structured interviews provide incremental validity over other selection devices. The findings are mixed as to whether a face-to-face structured interview adds to the prediction of job performance above what can be achieved with cognitive ability and personality tests (Cortina, Goldstein, Payne, Davison, & Gilliland, 2000; Walters, Miller, & Ree, 1993) or with a written version of the structured interview administered as a test (Whetzel, Baranowski, Petro, Curtin, & Fisher, 2003). Another unresolved issue is what accounts for the individual differences among interviewers in the validity of their judgments. Substantial variations have been found among interviewers in terms of the validity of their interview judgments (Van Iddekinge, Sager, Burnfield, & Heffner, 2006). The research that would identify the source of these individual differences has yet to be conducted.

The construct validity of structured interviews also remains uncertain. In terms of discriminant validity, there is little evidence that interviewers differentiate in their ratings among the various constructs that are intended to constitute the dimensions of the rating scale (Darr & Catano, 2008; Dipboye, 1992). In terms of convergent validity, it is unclear what constructs are actually measured by structured interviews. For instance, interviews do not appear to correlate highly with personality attributes as measured by self-report inventories (Roth, Van Iddekinge, Huffcutt, Eidson, & Schmit, 2005) and are only weakly related to scores on cognitive ability tests (Berry, Sackett, & Landers, 2007; Huffcut, Roth, & McDaniel, 1996; Salgado & Moscoso, 2002). Perhaps the most important remaining issue is that the existing knowledge base does not allow an identification of the relative importance of the elements of interview structure. In the meta-analyses of interview validity, each of the elements of interview structure (e.g., nature of questions) is inevitably confounded with other factors (e.g., rating scales used, training of interviewers). Research is needed that validates the separate components of structure and provides guidance to practitioners on how to assemble interview procedures that attain the highest validity at the lowest cost while remaining legal and fair.

SOCIAL PROCESS PERSPECTIVE ON THE SELECTION INTERVIEW

So far we have discussed the psychometric evaluations of interviewer judgments as though the interview were just another test or inventory. Although the interview can be evaluated on this basis, and previous validation research has yielded important insights, interviews are unique in several respects when compared with other selection instruments. For one, the interview is not limited to any particular set of constructs. Mental ability, personality, motivation, work-related knowledge and skills, and a variety of other constructs can become the focus of an interview and can be measured on the basis of interviewer questions. Even when the constructs measured in a selection interview are well defined and held constant, one cannot have the same degree of confidence that there is equivalency across situations as one can with a mental ability test or personality inventory because all judgments are subject to the interviewer's biases and processing limitations (Dipboye, 1992). Second, even the most structured interview formats provide an opportunity for deviations from what was intended in the procedures (Latham & Saari, 1984). These deviations are not as likely with paper-and-pencil selection instruments. For instance, it is highly unlikely that new questions will be inserted or existing questions deleted across administrations of a mental ability test or a personality inventory. Third, the measurement instrument in the interview is the individual interviewer, who is subject to all the limitations and frailties so often documented in the research on decision making and judgment. Moreover, the interviewer is typically tasked with not only assessing the applicant but also recruiting and providing information. The quality of assessments is likely influenced by how well interviewers can balance these sometimes competing objectives.

Above all, the interview is not just a score representing a final judgment by the interviewer. It is also a social interaction (Dipboye, 1982, 1992; Herriot, 2002). Interviews, by definition, consist of a

communication between a representative of the employer and the individual who seeks employment. This communication contains verbal and nonverbal acts on the part of both interviewer and applicant. Moreover, the interview is a dyadic interaction in which there is reciprocal influence. Even in a structured interview, in which there is an attempt to constrain the interaction, interviewers can convey nonverbal and paralinguistic behaviors that affect applicants. Likewise, applicant behavior can trigger interviewer responses. In its most idiosyncratic form, the interview is defined by the unique pairing of interviewer and applicant.

The uniqueness of the interview as an assessment tool requires that psychologists move beyond the examination of the reliability and validity of the final judgment. What is needed is an understanding of social process in the interview and the relationships between this process and the quality of the interviewer's assessments of the applicant (Dipboye, 1992). Thus, in the remainder of this chapter, we explore the research on the selection interview pertinent to three phases of this social process, depicted in Figure 27.1, and then theorize about the potential linkages between the process and the final psychometric outcomes of the interview.

First Phase: Encounter With Applicant Before and Early in the Interview

The interview process begins before the actual face-to-face conversation between interviewer and applicant. Interviewers have expectations of applicants based on the applicant pool and other information available to them before the session. The first encounter with the specific applicant is usually in a preview of paper credentials about the applicant such as unscored applications; scores on personality, ability, and knowledge tests; transcripts; references; the applicant's cover letter; and, in some cases, a portfolio containing examples of the applicant's prior work. Indeed, interviewers' preinterview evaluations of applicants are affected by paper credentials (Arnulf, Tegner, & Larssen, 2010; Brown & Campion, 1994; Cable & Gilovich, 1998; Cole, Feild, Giles, & Harris, 2004; Cole, Rubin, Feild, & Giles, 2007; Dalessio & Silverhart, 1994; Wade & Kinicki, 1997). Moreover, preinterview impressions are related to postinterview impressions at a level that is substantial and of potential practical importance. For example, Macan and Dipboye (1990) reviewed the research and reported a correlation of .35.

The information available to interviewers in the preinterview phase can result in valid predictions of performance. Schmidt and Hunter (1998) reported validities, corrected for unreliability in the criterion and range restriction on the predictor, of .51 for mental ability, .35 for scored biographical inventories, .26 for quantitative reference checks, .18 for years of job experience, .13 for training and experience ratings, and .11 for academic achievement. If interviewers fully use each of these items of information and optimally combine them in their preinterview and postinterview judgments, one could expect that the preview of credentials would lead to judgments that provide valid predictions of future performance. Of course, the clinical judgment research has suggested that interviewers will fail to achieve the validities that could be obtained with mechanical combinations of the information they gather before and during the interview (Dawes, Faust, & Meehl, 1989).

The initial encounter between the interviewer and applicant provides an additional source of information. Interviewers form impressions on the basis of appearance as well as the applicant's verbal and nonverbal behaviors during the introductions that occur at the beginning of the interview. Likewise,

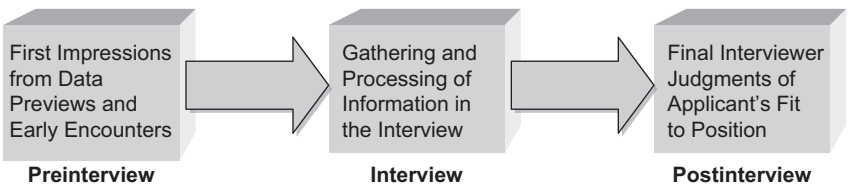
First Impressions from Data Previews and Early Encounters
Gathering and Processing of Information in the Interview
Final Interviewer Judgments of Applicant's Fit to Position
Preinterview
Interview
Postinterview

FIGURE 27.1. Three phases of the selection interview process.

research has suggested that the applicant's handshake can influence initial interview impressions (Stewart, Dustin, Barrick, & Darnold, 2008). Yet even before the handshake, interviewers are already forming impressions of job applicants on the basis of such factors as scent (Baron, 1983). In just seconds, individuals form judgments about others on the basis of their appearance and nonverbal behavior. Ambady and Rosenthal (1993) found that individuals' initial judgments of faculty on the basis of thin slices of nonverbal behavior as brief as 6, 15, and 30 seconds predicted future teaching ratings. Nonverbal behavior and physical appearance have been linked to judgments of applicants in both laboratory and field research (Barrick, Shaffer, & DeGrassi, 2009). That individuals form impressions so quickly is problematic given that early impressions can have a powerful effect on later judgments but show limited validity as predictors of later job performance (Barrick et al., 2009; Borman, 1982).

Social Interaction Between Interviewer and Applicant

In determining whether people obtain the work they desire, and in setting the stage for how their careers unfold, interviews are among the more important events in people's work life. Yet, the conversation that takes place between interviewer and applicant is typically a short encounter between strangers that is far removed from what people often consider an important relationship, such as friendships, romantic affairs, or cohesive work groups. Wish, Deutsch, and Kaplan (1976) identified four dimensions on the basis of similarity ratings of various types of interpersonal relationships: equal–unequal, competitive–cooperative, socioemotional–task oriented and formal, and superficial–intense. "Interviewer and job applicant" was located in this four-dimensional space at a point characterized by unequal and task-oriented relationships and was somewhat competitive and superficial in nature. The relationships that came closest to occupying the same point in this four-dimensional space were "master and servant" and "supervisor and employee."

What happens in the interview? The few studies that have attempted to account for the interactions between interviewer and applicant have provided a description that fits the Wish et al. (1976) profile (C. Anderson, 1960; Chapman & Zweig, 2005; Daniels & Otis, 1950; Silvester & Anderson, 2003; Stevens, 1998; Tengler & Jablin, 1983; Tullar, 1989). The typical interview is relatively short, lasting 30 minutes to an hour at most. During this time, the interviewer asks about 20 to 30 questions, approximately half of which are open-ended. Consistent with the expectation that the interviewer is in charge, when interviewers spend more time talking and less time asking questions, they tend to rate applicants more favorably and are more likely to recommend hiring than when they spend less time talking. The total utterances on the part of the interviewer and applicant are approximately equal, but a very small proportion of the interview session is typically devoted to questioning by the applicant. There is even some indication that applicants who ask questions, particularly about the interview process, are not perceived favorably (Babbitt & Jablin, 1985). Applicants who are most favorably perceived complement the behavior of the interviewer. They respond to dominance and structuring by the interviewer with submission and low structuring (Tullar, 1989).

Questioning by the interviewer. In terms of structure, one can distinguish between open-ended questions and closed-ended questions and between primary questions that begin the discussion of a topic and secondary questions that follow up on or attempt to flesh out responses to primary questions. As far as the content of the question, there is no limit to the topics that could be the focus of an interview question. Among the most frequent topics for questions in employment interviews are biographical information and job knowledge and skills. Less frequently assessed topics relate to values, opinions, and personality, which may be quite relevant if the objective is to assess person–organization fit (Kristof-Brown, 2000; Van Iddekinge, Raymark, & Roth, 2005).

Two types of questions that are part of structured interviewing formats are the situational- and the behavior-description question. Critical incident methodology is used in the development of both types of questions. This methodology involves the

generation of a particularly positive or negative event that occurred in the workplace (see Lance & Wilson, 1997, for a review of the technique). Incumbents, supervisors, and other subject matter experts relate actual cases of successful and unsuccessful performance in the job, and these incidents provide the basis for identifying criterion dimensions and the specific questions that are asked. Situational questions consist of hypothetical situations, derived from the critical incident analysis, that ask what applicants would do if they confronted the situation. For instance, if it were important to be able to work under time pressure on several tasks at the same time, one might pose a hypothetical situation involving competing demands and pressures to meet deadlines. The situation would typically be phrased to resemble the types of pressures and work demands faced in the position for which the applicant is being interviewed.

In behavioral description questions, applicants are asked to describe what they did do in various types of situations in the past. For instance, if the ability to plan and organize work under a heavy workload were an important factor to consider, recent graduates with little work experience might be asked, "What did you do when you were a student when you were faced with several assignments all due at the same time?" The most comprehensive meta-analysis conducted comparing these two types of questions concluded that behavioral description interview questions were somewhat superior to situational questions (Taylor & Small, 2002). However, a variety of other factors confound the comparisons of these two types of questions, and both remain viable alternatives. For example, situational interviews might be more useful for applicants who have no prior work experience and, therefore, would be unable to answer a behavioral description question.

Impression management by the applicant. Much of the social interaction that takes place during the interview session can be described as impression management by the applicant, which can be defined as attempts of the applicant to convey a desired image and convince the interviewer that this is an accurate description of his or her characteristics (e.g., Ellis, West, Ryan, & DeShon, 2002; Peeters & Lievens, 2006; Van Iddekinge, McFarland, & Raymark, 2007). A common distinction is between assertive impression management tactics that are oriented to conveying a positive image and defensive tactics that are aimed at defending against and preventing negative impressions. Assertive tactics are the most common type of impression management in selection interviews. Of the assertive tactics, self-promotion is the most frequently used (Stevens & Kristof, 1995). Here, the applicant attempts to enhance the impression that is conveyed to the interviewer through the use of explicit self-descriptions, claiming responsibility for past positive events, embellishing the importance of the positive events for which they claim responsibility, and telling stories of overcoming difficulties to achieve success. One implication is that the applicant's statements and behavior in interviews are performances, not merely responses to be taken at face value. It should come as no surprise that applicants attempt to make a good impression. The extent to which self-promotion damages the validity of interviewer judgments is debatable and in need of further research. One view is that any impression management is tantamount to lying to be ferreted out by the interviewer (Levashina & Campion, 2007). We agree with the more benign view of applicant impression management as a natural occurrence that can occur as outright lies but more frequently occurs as the applicant's attempts to negotiate between telling the truth while still conveying a positive impression (Marcus, 2009). From this perspective, impression management can be a legitimate and natural behavior that does not necessarily have an adverse effect on criterion-related validity.

The good interview script. Both interviewer and applicant appear to have cognitive scripts for how the interview will unfold (Stevens, 1998), and interviews are perceived as having gone well or not well on the basis of how well the session conforms to this script. The typical interviewing advice fits this script of the good interview. Interviewers are advised to put applicants at ease and to provide verbal and nonverbal reinforcement for the applicant's disclosure of information relevant to evaluating competencies and to maintain objectivity and avoid leading

questions (Janz, Hellervik, & Gilmore, 1986). Consistent with this advice, applicants respond more positively to interviewers who conduct the interview session in a manner that conveys that they are warm and friendly, knowledgeable about the job, and competent (Chapman, Uggerslev, Carroll, Piasentin, & Jones, 2005; Derous, 2007; Schreurs et al., 2005). Likewise, interviewers evaluate the sessions they have conducted as having gone well to the extent that they established rapport with the applicant (Chapman & Zweig, 2005).

The research on rapport has shown that an interaction that goes well is marked by accommodation on the dominance–submission dimension and convergence on the warmth–affection dimension (Markey, Funder, & Ozer, 2003; Sadler, Ethier, Gunn, Duong, & Woody, 2009). Generalizing from this research to the interview, we would hypothesize that when the interview goes well, a complementarity occurs in which one party tends to respond with dominance to the other's submission and with submission to the other's dominance, rather than matching dominance with dominance and submission with submission (Tullar, 1989). We would also hypothesize that a matching of nonverbal and verbal behavior occurs in which interviewer and applicant imitate the other's actions (Matarazzo & Wiens, 1972).

It would appear from this research that the good interview is one in which the interviewer and applicant are in synch. One consequence that has been demonstrated in other types of social interactions and may well apply to the selection interview is *emotional contagion*, which is the unconscious transfer of emotions between people that occurs when individuals mimic and synchronize the facial expressions, vocalizations, and movements of others (Hatfield, Cacioppo, & Rapson, 1992). When one imitates these nonverbal and paralinguistic behaviors, a process called *motor mimicry* (Chartrand & Bargh, 1999), the individual will begin to experience the emotion that he or she is mimicking. Indeed, simply exhibiting a facial expression can elicit the corresponding emotion in an individual (Howard & Gengler, 2001; Larsen & Kasimatis, 1990).

The power of emotional contagion rests in its automaticity and unconscious nature (Johnson & Johnson, 2009). For example, emotional contagion can occur within a 2-minute silent interaction (Friedman & Riggio, 1981) and can occur without individuals even recognizing the change in their emotions (Neumann & Strack, 2000), making it difficult to avoid the contagion process. Although this effect has moderators, such as the emotion sender's expressiveness (e.g., Sullins, 1991) and the receiver's susceptibility to emotional contagion (e.g., Johnson, 2008), it appears that this effect is widely pervasive. Although emotional contagion has not been examined in the selection interview, we expect that emotional contagion is likely to occur in interview contexts and, depending on the positivity of the emotions conveyed, determines whether the session is perceived as having gone well or not.

It is possible that a job applicant who expresses positive emotion could cause the interviewer to catch those positive emotions. Likewise, an interviewer who is expressing negative emotions might spread those emotions to the job applicant. This direction of emotional contagion (from interviewer to applicant) is more likely to occur given that people of higher status are more likely to send their emotions to people of lower status than the other way around (e.g., Polansky, Lippitt, & Redl, 1950). More important, as demonstrated in other domains, the resulting emotions have the possibility to affect the applicant's (or interviewer's) attitudes or performance (e.g., Johnson, 2008, 2009). More specifically, the resulting emotions could enhance an interviewer's perceptions of an applicant's employment suitability or interview performance just as it could affect the applicant's impressions of the interviewer or the organization.

In conclusion, interviewers and applicants appear to react positively to their time together to the extent that the sessions are characterized by rapport and positive emotions (Chapman & Zweig, 2005; Chapman et al., 2005; Derous, 2007; Schreurs et al., 2005). Structuring the interview process may prevent such rapport and lead to slippage in the implementation of these procedures (Dipboye, 1994). The strong desire for face-to-face contact may also account for the negative reactions to videoconferencing that have been found. For example, applicants perceive interviewers as less friendly in videoconference interviews than in face-to-face

interviews or phone interviews (Straus, Miles, & Levesque, 2001). Likewise, compared with videoconference or phone interviews, applicants perceive face-to-face interviews as fairer and report greater intentions of accepting a job offer (Chapman, Uggerslev, & Webster, 2003). Litigation intentions are also lower in face-to-face interviews than in videoconference or telephone interviews (Bauer, Truxillo, Paronto, Weekley, & Campion, 2004).

INTERVIEWER'S POSTINTERVIEW JUDGMENTS OF APPLICANT

Before, during, and after the interview, the interviewer forms an impression of the applicant's fit to the position, culminating in a final judgment of whether the applicant should be hired.

Dual process theories depict the information processing in situations such as selection interviews as shifting between the categorical and the individuating and the conscious and the unconscious (Fiske, Lin, & Neuberg, 1999). Thus, the first reaction might be an unconscious and automatic categorization of the applicant. If information is found that contradicts the initial impression, the interviewer might rely on a subtype. For example, the interviewer might go from categorizing an applicant as a typical woman to subtyping her as typical professional woman. To the extent that information contradicts a type or subtype, the interviewer will shift to a deliberate, conscious, and piecemeal gathering and integration of information. A large amount of research has addressed how various factors influence the processing of information from the interview and the postinterview judgments of applicant fit that we review next.

Applicant Qualifications and Style

Research on the effect of applicants' objective qualifications on interviewers' postinterview judgments has shown the importance of objective information such as grades, work experience, test scores, and biodata (Graves & Powell, 1988; Huffcutt, Roth, & McDaniel, 1996; Singer & Bruhns, 1991). However, the applicant's self-presentation style appears to be as important as the objective qualifications. More positive evaluations are given to applicants who act verbally and nonverbally in a manner that conveys enthusiasm, interest in the job, self-confidence, and positivity (N. Anderson, 1991; N. Anderson & Shackleton, 1990; Einhorn, 1981). Interviewers evaluate applicants more positively to the extent that they convey high immediacy of nonverbal behavior in the form of smiling, eye contact, and forward lean in body orientation (Barrick et al., 2009; Levine & Feldman, 2002); vocal attractiveness and fluency (DeGroot & Kluemper, 2007); physical attractiveness (Barrick et al., 2009; Hosoda, Stone-Romero, & Coats, 2003); and assertive impression management tactics (Barrick et al., 2009). Interviewers also give more positive evaluations to applicants who speak with grammatically correct English and standard dialect and avoid paralinguistic behaviors such "you know," "uh," and "um" (Hollandsworth, Kazelskis, Stevens, & Dressel, 1979; Russell, Perkins, & Grinnell, 2008).

Although often considered to constitute an irrelevant basis for evaluating applicants, some recent research has suggested that interviewers can infer from nonverbal behavior personality traits that are meaningfully related to job performance (DeGroot & Gooty, 2009). Also, a modest correlation has been found between observer ratings of paralinguistic and nonverbal interview and job performance (Motowidlo & Burnett, 1995). Despite these individual studies, the findings from a meta-analysis have suggested that applicant nonverbal behavior, appearance, and impression management are unrelated to job performance (Barrick et al., 2009).

Applicant Appearance, Demographic Characteristics, and Sexual Orientation

With the passage of civil rights legislation, it is illegal in most cases to base evaluations on applicant race, ethnicity, nationality, gender, disability status, and age. Physical attractiveness, obesity, and sexual orientation are not banned as bases of interviewer judgments but are generally seen as irrelevant and inappropriate bases for evaluating qualifications for most positions. Considerable attention has been given to assessing the extent to which these factors play a part in interviewers' evaluations of applicant qualifications.

Race, ethnicity, and nationality. At least two qualitative reviews of the field research have concluded that little adverse impact on basis of race and ethnicity occurs as the result of using interviews (McCarthy, Van Iddekinge, & Campion, 2010; Morgeson, Reider, Campion, & Bull, 2008; Posthuma, Morgeson, & Campion, 2002). Roth et al. (2002) questioned these conclusions and suggested that past research has underestimated the magnitude of ethnic group differences as a result of failure to correct for range restriction.

Gender. Several qualitative reviews of the field research have concluded that little adverse impact has occurred on basis of gender as the result of using interviews (McCarthy et al., 2010; Morgeson et al., 2008; Posthuma et al., 2002). In another qualitative review, Graves (1999) concluded that gender bias is more likely to occur with repeated-measures designs and low amounts of information provided on the applicant. Davison and Burke (2000) concluded from their meta-analysis that women receive lower ratings than men for male-typed jobs and men receive lower ratings than women for female-typed jobs.

Age. Morgeson et al. (2008) concluded from a qualitative review of both field and laboratory research that older applicants receive less favorable ratings than younger applicants in laboratory research. However, they questioned the generalizability of the laboratory research and suggested that less age discrimination occurs in the field. Whether their conclusions are correct and age bias is limited to the lab is conjecture and remains to be tested.

Physical attractiveness and obesity. The evidence that raters in the role of interviewers are biased against applicants who are less physically attractive is strong and consistent (Barrick et al., 2009; Hosoda et al., 2003). Moreover, Barrick et al. (2009) found that the effect is as strong in the field as it is in the laboratory. Similar to the findings for attractiveness, Rudolph, Wells, Weller, and Baltes (2009) found in a meta-analysis a strong and consistent bias against overweight applicants.

Disability. The research on bias against people with disabilities is mixed and inconclusive (Colella & Stone, 2005). The effect of applicant disability on interviewer evaluations depends on the nature of the disability. For example, it appears that applicants with mental health or substance abuse disabilities are treated with much less sympathy than applicants with physical disabilities such as cancer (Reilly, Bocketti, Maser, & Wennet, 2006). In the case of invisible disabilities, the findings are mixed as to whether disclosure to the interviewer will lead to positive or negative evaluations (Dalgin & Bellini, 2008; Roberts & Macan, 2006).

Sexual orientation. We could find no research on discrimination on the basis of sexual orientation in the interview except for a field study conducted by Hebl, Foster, Mannix, and Dovidio (2002). In this study, confederates sought employment in retail stores wearing a baseball cap on which was printed either the slogan "Gay and Proud" or "Texan and Proud." Those confederates with the gay-and-proud cap received fewer job offers than those wearing the alternative cap. Although the difference was statistically nonsignificant and sexual orientation is not protected under federal law, this finding would constitute adverse impact under the four-fifths rule of the Equal Employment Opportunity Commission.

Subtle Biases in Interviewer Evaluations and Conduct of the Session

The research we have reviewed thus far has examined the explicit ratings of applicants who were depicted in a transparent manner as varying on the basis of illegal, irrelevant, and inappropriate attributes. An increasing amount of research in social psychology has suggested that even when interviewers suppress their prejudices, unconscious biases can affect the outcomes of an interview.

Effects of subtle cues associating applicant with outgroup. An *outgroup* is any social group toward which one has a negative affective or behavioral response on the basis of differences in group membership. For example, one may feel negatively toward people of a different racial, religious, or socioeconomic group. Bias against outgroup applicants may be more likely to emerge as a result of subtle cues associated with outgroup membership than more transparent cues such as naming ethnicity or gender in the application or providing a

picture. Among the cues that appear to prime prejudicial attitudes and evaluations, possibly without the evaluator's awareness, are accent (Hosoda & Stone-Romero, 2010; Purkiss, Perrewé, Gillespie, Mayes, & Ferris, 2006), name (King, Madera, Hebl, Knight, & Mendoza, 2006; Purkiss et al., 2006), scent (Sczesny & Stahlberg, 2002), and physical attractiveness (Johnson, Podratz, Dipboye, & Gibbons, 2010).

Available justifications for bias. If a socially acceptable justification exists for the negative evaluation of the outgroup member, implicit biases can emerge and shape explicit evaluations. One possible demonstration of this comes from a field experiment in which applicants for low-paid jobs submitted résumés that were equivalent with the exception of a prison record (Pager, Western, & Sugie, 2009). The researchers found that when the applicant was White and had a prison record, 22% were invited for an interview, but when the applicant was Black and had a prison record, only 10% were invited. One interpretation is that prejudice against Black applicants, combined with a convenient and socially acceptable justification for rejection, led to the higher rejection of Black applicants. Another demonstration comes from a laboratory simulation (Ziegert & Hanges, 2005). To the extent that the student participants held implicit biases against Blacks and the climate or the organization supported discrimination, the participants were more likely to discriminate against Black candidates. According to Ziegert and Hanges (2005), a prejudicial climate in which higher management expressed a preference for a White candidate provided the justification for a discriminatory act that was suppressed in the absence of justification.

Leakage of implicit attitudes into the conduct of the interview. Interviewer biases may not be manifested in the external evaluations of applicants, which they can control, as much as in the nonverbal and paralinguistic behaviors, which they cannot control. Word, Zanna, and Cooper (1974) had White students interview Black and White students in simulated job interviews. The White interviewers were shown to exhibit a variety of negative nonverbal behaviors in interviewing Black applicants. White interviewers interviewing Black applicants, compared with those interviewing White applicants, spent 25% less time interviewing, seated themselves more distantly from the applicant, and displayed nonverbal behavior that reflected less openness and responsiveness (e.g., less eye contact, forward lean, shoulder orientation). Hebl et al. (2002) provided a more recent field demonstration of how biases can leak into treatment of job applicants who are identified as gay.

Caveats. Laboratory investigations intentionally create artificial environments so as to test theory. They are intended to assess whether discrimination can occur and should not be used to determine the frequency or magnitude with which discrimination does occur outside the laboratory (Mook, 1983). Consequently, it is misleading to conclude from either a qualitative or a quantitative review of laboratory research on interviewer bias that such bias is strong, weak, or nonexistent in the field. A limitation that we would suggest is just as serious is the file-drawer effect in the field research on discrimination. We suspect that few for-profit, private organizations would agree to the publication of the results of a study showing bias. As a consequence, we suspect that there are serious limitations to generalizing the findings of research in which the organization's permission to publish is required. In our opinion, unobtrusive field research provides a better basis for judging the prevalence or strength of discrimination (e.g., Agerstrom & Rooth, 2011; Hebl et al., 2002; King, Shapiro, Hebl, Singletary, & Turner, 2006; Newman & Krzystofiak, 1979; Pager, Western, & Bonikowski, 2009; Pager, Western, & Sugie, 2009).

LINKAGES BETWEEN IMPRESSIONS AND INTERVIEWER INFORMATION PROCESSING AND GATHERING

Each of the phases that we have described here occurs in the context of the preceding phase. Moreover, each separate phase could be described as a series of episodes, each of which occurs in the context of the preceding episode. The interviewer's conduct of the interview and the processing of information gathered are linked to the interviewer's prior impression of the applicant. On the basis of

previous theoretical and empirical investigations of the effects of prior expectancies on impression formation (Kruglanski & Mayseless, 1988; Snyder & Klein, 2005; Trope & Bassok, 1982), three types of linkages among the phases can be hypothesized: diagnostic, confirmatory, and disconfirmatory. In contrasting these three processes, let us assume that an interviewer begins the interview with the impression of how well the applicant's qualifications provide a good fit to the job. The three alternative linkages describe how this initial impression might influence the subsequent gathering and processing of information.

Diagnostic Linkage

In the diagnostic linkage, the initial impression is a two-sided hypothesis that the interviewer attempts to test in a scientific fashion by searching for and giving preference to information that is diagnostic in the evaluation of the applicant's fit to the position. We follow Trope and Bassok's (1982) conceptualization of diagnosticity. The diagnosticity of an item of information on an applicant (I) used in assessing a job-related attribute (A) is a function of two conditional probabilities: (a) the probability that A is true for the applicant if the item of information is descriptive of the applicant, and (b) the probability that attribute is true for the applicant if the item of information is not descriptive of the applicant. Objective diagnosticity of an item of information gathered in the interview increases as the difference between the two conditional probabilities increases. A diagnostic question would be one in which qualified and unqualified applicants differ in how they answer the question. If qualified applicants were much more likely than unqualified applicants to say that they like large parties in response to the question "Do you like large parties?" then that question would be diagnostic of qualifications. However, if qualified and unqualified applicants were both likely to answer this question in the same way, then the question would be low on diagnosticity. Prior impressions can engage a diagnostic process of information gathering and processing in which interviewers are guided by a search for information that delineates between applicants who possess the attribute that is being evaluated and those who do not.

Confirmatory Linkage

With the confirmatory linkage, the initial impression moves an interviewer in the direction of maintaining consistency with this impression. The consistency of information with the prior impression is defined by the probability that a job-related attribute (A) is true of an applicant given an item of information (I) that describes the applicant. Consistency of I with an impression is independent of diagnosticity in testing that impression. To use our previous example, interviewers holding a positive initial impression that an applicant fits a job requiring extraversion might tend to ask the question "Do you like large parties?" because it is consistent with the initial impression. As noted earlier, the question may or may not be diagnostic depending on whether qualified applicants answer differently than unqualified applicants. Nonetheless, a confirmatory linkage would suggest that interviewers who have a positive impression of the applicant's qualifications will ask this question because it seeks information that is consistent with the impression. If qualified and unqualified applicants differ in how they answer the question, the linkage may be diagnostic as well as confirmatory, but to the extent that qualified and unqualified applicants do not differ, the linkage is confirmatory but nondiagnostic. In some cases, a confirmatory process constitutes a self-fulfilling prophecy in which there is a behavioral confirmation of the interviewer's impressions (Dipboye, 1982; Dougherty, Turban, & Callender, 1994). For instance, if interviewers with a positive initial impression ask what the applicant likes about large parties, the applicant may well generate examples of things he or she likes about large parties. Because the question is slanted in the direction of probing for information that is consistent with the positive impression, the subsequent information gathering may tend to confirm the initial impression. Swann and Ely (1984) found in simulated job interviews that such self-fulfilling prophecies were most likely when interviewers were highly confident of their preinterview impressions of the applicant and the applicant lacked confidence.

Disconfirmatory Linkage

With a disconfirmatory linkage, the initial positive impression moves the interviewer in the direction

of seeking, attending to, and retaining information that is inconsistent with prior impressions. Thus, interviewers form a positive impression of an applicant's fit to the position and in their information gathering try to show that this impression is false, that is, that the applicant is unqualified for the position. Thus, interviewers would seek, retain, and give more weight to information that shows that the applicant is unqualified than they would to information that shows the applicant is qualified. Binning, Goldstein, Garcia, and Scattaregia (1988) and Neuberg (1989) found in simulated interviews that interviewers were biased in favor of applicants for whom they had negative preinterview impressions when prompted with the goal of accurate impression formation or when they questioned someone of the opposite sex.

Different Interviewer Motivations May Underlie the Same Linkage

Interviewers can approach the conduct of the interview with a variety of motives. One would hope that the motivation to accurately evaluate the applicant and pick the right person for the job is the dominant motive, but other potential motives are at work in hiring situations. For one, interviewers may be motivated to manage impressions, such as conveying the image of fairness and freedom from bias or evaluative rigor. They may also be motivated to win the approval or avoid the disapproval of supervisors, peers, and other potential observers of the interview outcome. Each of the alternative linkages described in the preceding sections can be motivated by a variety of motives, including the desire to make an accurate decision, convey a particular image, or win approval. Take, for example, an interviewer who has an initial negative impression of an applicant. Also assume that this interviewer seeks to avoid the appearance of being biased against women. In such a case, the interviewer might well seek evidence that the applicant is qualified and may pay special attention to such information (Binning et al., 1988). It is also possible that a desire to make an accurate decision may lead to confirmatory, disconfirmatory, or a diagnostic process. However, which process is most likely to lead to the most valid judgments?

Effects on Criterion-Related and Construct Validity

We would hypothesize that diagnostic processes will benefit the quality of interviewer judgments to a greater extent than either confirmatory or disconfirmatory processes. Although this proposition appears reasonable, it is speculative, because few studies have been conducted to assess the relationship of the interview process to the psychometric quality of interviewer judgments. Also, we would not claim that valid judgments are never associated with confirmatory or disconfirmatory processes. A confirmatory process might allow for valid interviewer judgments in which prior impressions are based on valid indicators of future performance and the interviewer optimally uses and combines this information with other indicators of performance. Likewise, a disconfirmatory process might work in generating valid judgments in which an initial impression is erroneous and the interviewer searches for and processes information that challenges that impression. We should further note that confirmatory and disconfirmatory linkages are not necessarily driven by blatant irrationality but may reflect what interviewers perceive to be the more diagnostic approach to assessing applicants. Nevertheless, we would hypothesize that interviewer postinterview judgments will show higher criterion-related validities to the extent that information gathering and processing are diagnostic as opposed to confirmatory or disconfirmatory. In other words, more valid judgments follow from attempts by the interviewer to seek and retain information that allows him or her to distinguish applicants with higher qualifications for the job from those with lower qualifications. The question then becomes, how does one design the interview so that interviewers are likely to seek and retain diagnostic information rather than using purely confirmatory or disconfirmatory strategies?

STRUCTURING THE INTERVIEW TO INCREASE DIAGNOSTICITY

We would propose several interventions to enhance the diagnosticity of the interviewer's information gathering and processing. Several of our suggestions are consistent with what is typically thought to constitute a structured interview. Several other of

our suggestions deviate from what is typically described as a structured interview. We present a summary of these elements in Table 27.1.

Structure Questions So That They Are Diagnostic for the Criterion

One of the characteristics of behavioral description and situational interviews, which have been shown to achieve higher validities than unstructured interviews, is that questions are predetermined on the basis of their job relatedness. First is a critical incidents analysis in which incumbents relate cases of successful and unsuccessful performance on the job and then questions are constructed to gather information that relates to these incidents and the criterion dimensions that they represent. To the extent that the questions asked are structured to differentiate high-performing from low-performing employees, they are by definition diagnostic. We would hypothesize that the criterion-related validity of interviewer judgments will increase as the frequency of such questions increases.

TABLE 27.1

Ways of Structuring the Interview to Increase Diagnosticity: Good and Bad Examples

Interview component	Good example	Bad example
Structure questions so they are diagnostic.	Use critical incidents technique to identify performance issues that are critical in distinguishing effective and ineffective employees in the specific position. Structure questions that focus on these critical areas and that research has shown distinguish between applicants who later succeed in the position and those who do not.	Ask questions that are typically asked in interviews regardless of their relevance to the position for which the applicant is being considered. For example, "What are your hobbies?" "What are your greatest strengths and weaknesses?" "What did you like and dislike about your previous job?"
Require consistency in asking questions.	Interviewers are required to ask the same questions of all applicants and to ask them in the same order, using the same wording.	Interviewers are allowed to ask whatever set of questions seem to make sense for the specific applicant being interviewed in whatever order they wish.
Increase diagnosticity of information processing.	Once questions are generated that focus on critical areas of performance that distinguish effective and less effective employees, conduct research to identify the specific answers that distinguish effective from less effective employees. Link alternative answers to numerical scores that the interviewer can use in rating applicants.	Interviewers form holistic judgments of applicant qualifications based on their gut feelings about how well each applicant answered.
Turn early impressions into hypotheses to be tested.	Either have interviewers avoid forming initial impressions or else emphasize that they should treat these as tentative hypotheses to be tested in the interview session.	Encourage interviewers to form a quick impression and emphasize that the first impression is usually the best impression.
Reduce cognitive load of interviewer.	Limit the task of the interviewer to gathering information on job qualifications and evaluating these qualifications.	Make the interviewer responsible for gathering information on job qualifications, evaluating these qualifications, recruiting the best applicants, counseling the applicants on their job search, and providing information to the applicant on the nature of the job and organization.
Increase rapport between interviewer and applicant.	Provide explanation of the interview procedure to the applicant; allow the applicant the opportunity to ask questions of the interviewer at some point in the session; begin the interview with casual chit-chat.	Conduct an interrogation in which applicants are not allowed to ask questions, interviewers are not allowed any informal exchanges with applicants, and no explanation is given to applicants of the interview procedures.

Require Consistency in the Asking of Questions Across Applicants

Another key component of behavioral description and situational interviews is that all interviewers must adhere to the same line of questioning. In the most highly structured interviews, interviewers are not even allowed to ask follow-up questions or to probe previous answers. The research that would allow one to determine just how much consistency is required or whether interviews should become question–answer interrogations has not been conducted. It appears safe to conclude, however, that requiring some consistency is more likely to lead to diagnostic interviewing than is achieved in interviews in which no consistency requirement is imposed. We would suggest that while maintaining consistency in the content and order of questions asked across applicants, interviewers should be allowed to use follow-up probes and secondary questions. We base this on evidence that little validity is added at the highest levels of structure above the validities obtained with interviews that allow more discretion in questioning (Huffcutt & Woehr, 1999).

Increase Diagnosticity of Information Processing by Linking Answers to Rating Options

The use of behaviorally anchored rating scales appears to lead to more accurate judgments of interviews than graphic rating scales lacking behavioral anchors (Maurer, 2002). Higher criterion-related validities have been shown as a consequence of behavioral rating scales and may be more important than the structure of the interview questions (Taylor & Small, 2002). We would suggest that when behaviorally anchored rating scales are used, interviewers approach their use of information gleaned in the session in a more diagnostic fashion by mapping what they hear onto the crucial requirements of the position.

Turn Early Impressions Into Hypotheses to Be Tested

One other recommendation that is made in the most highly structured interview formats is to avoid previewing the applicant's paper credentials. We suspect that this aspect of interview structure is commonly violated in that most interviewers expect to see paper credentials before the interview (Chapman & Zweig, 2005). Indeed, Janz et al. (1986) included previewing applicant credentials as part of the procedures included in the behavior description interview. Assuming that most interviewers, even in structured formats, will have available some prior information on the applicant, the question that arises is the type of impression that interviewers should form from these credentials. We would speculate on the basis of previous research in social cognition (Kruglanski & Mayseless, 1988) that interviewers should be encouraged to form hypotheses rather than impressions and that these hypotheses should be two sided rather than one sided.

Reduce Cognitive Load

One reason that structured interviews may achieve more valid judgments than unstructured interviews is that they reduce the cognitive load on the interviewer. The typical interview requires too much of interviewers. Not only must the interviewer generate questions, observe the verbal and nonverbal responses to these questions, and retrieve this information to form a judgment, but they are also tasked with recruiting, counseling, and serving as a representative for the organization. All of this must happen within a short session, often 30 minutes or shorter. One clear finding from the social cognitive and cognitive research literatures is that deadlines and heavy cognitive workloads are likely to lead to shortcuts (Biesanz, Neuberg, Smith, Asher, & Judice, 2001; Freund, Kruglanski, & Shpitzajzen, 1985; Kruglanski & Mayseless, 1988; Nordstrom, Hall, & Bartels, 1998; Sherman & Frost, 2000). We would speculate that confirmatory and disconfirmatory processes are most likely to emerge and diagnostic processes are least likely to emerge when interviewers are faced with a high cognitive workload. As such, structuring the interview, and taking other measures to reduce cognitive load for the interviewer, have the potential to improve the interviewer's ability to engage in diagnostic processing.

More Rapport Leads to Better Information and Judgments

Previously, we suggested that interviews characterized by reciprocity in the interaction between

interviewer and applicant generate higher quality and quantity of information. Surprisingly little research has tested this notion and none that we could find was conducted in the context of selection interviews. Powell and O'Neal (1976) found that student observers of an interaction made more accurate, differentiated, and confident judgments of an interviewee's personality traits when the communication was characterized by reciprocity rather than being one directional, such as would occur in a structured interview. In a forensics context, Collins, Lincoln, and Frank (2002) found that interviewers who established rapport with interviewees produced substantially more correct items of information than interviewers who conducted the sessions with a harsh or neutral tone. The suggestion to increase rapport appears to be the opposite of what is suggested in the guidelines on structured interview. Yet, we would speculate that these elements will generate more diagnostic information that will improve the quality of interviewer judgments and more than compensate for the lower standardization in the process.

SUMMARY AND CONCLUSIONS

Despite the development of highly sophisticated tests that have proven worth in employee selection, the popularity of the face-to-face interview continues unabated. The interview remains the most frequently used means of assessment and is frequently the only tool used for selection. Even when other assessment tools are used, the information yielded from these measures is often filtered through the judgments of the interviewer. Structured interviews achieve higher validities and reliabilities than unstructured interviews, but little is known about the processes that account for these differences. We hypothesized that, to the extent that prior impressions lead to diagnostic information gathering and processing, interviewer judgments will demonstrate higher levels of criterion-related validity. We suggested several approaches to increase the diagnosticity of interviews, some of which are consistent with structured interview procedures and some of which are not. These approaches include using questions that differentiate high and low levels of employee performance; ensuring consistency in the conduct of the session across applicants; reducing interviewer cognitive load; using rating scales that map answers onto criterion dimensions; forming hypotheses rather than conclusions on the basis of preinterview information related to applicant qualifications; and establishing rapport with applicants during the interview process.

A basic assumption of this chapter is that a more conceptual approach to the interview is needed to guide research in both the field and the lab. The insight that is gained from theory and research will provide the basis for a more complete understanding of the successes and failures of interviewer judgment as a means of assessing applicant potential. Moreover, given the pervasive use of the interview, and its centrality to the selection process, such understanding will enhance the use of other tools of assessment that are so often intertwined with the interview.

References

Agerström, J., & Rooth, D. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology, 96*, 790–805. doi:10.1037/a0021594

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*, 431–441. doi:10.1037/0022-3514.64.3.431

Anderson, C. (1960). The relation between speaking times and decision in the employment interview. *Journal of Applied Psychology, 44*, 267–268. doi:10.1037/h0042783

Anderson, N. (1991). Decision making in the graduate selection interview: An experimental investigation. *Human Relations, 44*, 403–417. doi:10.1177/001872679104400407

Anderson, N., & Shackleton, V. (1990). Decision making in the graduate selection interview: A field study. *Journal of Occupational Psychology, 63*, 63–76. doi:10.1111/j.2044-8325.1990.tb00510.x

Arnulf, J., Tegner, L., & Larssen, Ø. (2010). Impression making by résumé layout: Its impact on the probability of being shortlisted. *European Journal of Work and Organizational Psychology, 19*, 221–230. doi:10.1080/13594320902903613

Babbitt, L. V., & Jablin, F. M. (1985). Characteristics of applicants' questions and employment screening interview outcomes. *Human Communication Research, 11*, 507–535. doi:10.1111/j.1468-2958.1985.tb00058.x

Baron, R. A. (1983). "Sweet smell of success"? The impact of pleasant artificial scents on evaluations of job applicants. *Journal of Applied Psychology, 68*, 709–713. doi:10.1037/0021-9010.68.4.709

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology, 94*, 1394–1411. doi:10.1037/a0016532

Bauer, T. N., Truxillo, D. M., Paronto, M. E., Weekley, J. A., & Campion, M. A. (2004). Applicant reactions to different selection technology: Face-to-face, interactive voice response, and computer-assisted telephone screening interviews. *International Journal of Selection and Assessment, 12*, 135–148. doi:10.1111/j.0965-075X.2004.00269.x

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview–cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology, 60*, 837–874. doi:10.1111/j.1744-6570.2007.00093.x

Biesanz, J., Neuberg, S., Smith, D., Asher, T., & Judice, T. (2001). When accuracy-motivated perceivers fail: Limited attentional resources and the reemerging self-fulfilling prophecy. *Personality and Social Psychology Bulletin, 27*, 621–629. doi:10.1177/0146167201275010

Binning, J. F., Goldstein, M. A., Garcia, M. F., & Scattaregia, J. H. (1988). Effects of preinterview impressions on questioning strategies in same- and opposite-sex employment interviews. *Journal of Applied Psychology, 73*, 30–37. doi:10.1037/0021-9010.73.1.30

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67*, 3–9. doi:10.1037/0021-9010.67.1.3

Brown, B., & Campion, M. (1994). Biodata phenomenology: Recruiters' perceptions and use of biographical information in resume screening. *Journal of Applied Psychology, 79*, 897–908. doi:10.1037/0021-9010.79.6.897

Cable, D., & Gilovich, T. (1998). Looked over or overlooked? Prescreening decisions and postinterview evaluations. *Journal of Applied Psychology, 83*, 501–508. doi:10.1037/0021-9010.83.3.501

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*, 655–702. doi:10.1111/j.1744-6570.1997.tb00709.x

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41,* 25–42. doi:10.1111/j.1744-6570.1988.tb00630.x

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology, 90*, 928–944. doi:10.1037/0021-9010.90.5.928

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Journal of Applied Psychology, 88*, 944–953. doi:10.1037/0021-9010.88.5.944

Chapman, D. S., & Zweig, D. I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personnel Psychology, 58*, 673–702. doi:10.1111/j.1744-6570.2005.00516.x

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893–910. doi:10.1037/0022-3514.76.6.893

Cole, M., Rubin, R., Feild, H., & Giles, W. (2007). Recruiters' perceptions and use of applicant résumé information: Screening the recent graduate. *Applied Psychology, 56*, 319–343. doi:10.1111/j.1464-0597.2007.00288.x

Cole, M. S., Feild, H. S., Giles, W. F., & Harris, S. G. (2004). Job type and recruiters' inferences of applicant personality drawn from resume biodata: Their relationships with hiring recommendations. *International Journal of Selection and Assessment, 12*, 363–367. doi:10.1111/j.0965-075X.2004.00291.x

Colella, A., & Stone, D. (2005). Workplace discrimination toward persons with disabilities: A call for some new research directions. In R. L. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 227–255). Mahwah, NJ: Erlbaum.

Collins, R., Lincoln, R., & Frank, M. G. (2002). The effect of rapport in forensic interviewing. *Psychiatry, Psychology and Law, 9*, 69–78. doi:10.1375/pplt.2002.9.1.69

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565–579. doi:10.1037/0021-9010.80.5.565

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*, 325–351. doi:10.1111/j.1744-6570.2000.tb00204.x

Dalessio, A., & Silverhart, T. (1994). Combining biodata test and interview information: Predicting decisions and performance criteria. *Personnel Psychology, 47*, 303–315. doi:10.1111/j.1744-6570.1994.tb01726.x

Dalgin, R. S., & Bellini, J. (2008). Invisible disability disclosure in an employment interview: Impact on employers' hiring decisions and views of employability. *Rehabilitation Counseling Bulletin, 52*, 6–15. doi:10.1177/0034355207311311

Daniels, H. W., & Otis, J. L. (1950). A method for analyzing employment interviews. *Personnel Psychology, 3*, 425–444. doi:10.1111/j.1744-6570.1950.tb01717.x

Darr, W., & Catano, V. M. (2008). Multi assessments of behavioral competencies and selection interview performance. *International Journal of Selection and Assessment, 16*, 68–72. doi:10.1111/j.1468-2389.2008.00410.x

Davison, H., & Burke, M. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior, 56*, 225–248. doi:10.1006/jvbe.1999.1711

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674. doi:10.1126/science.2648573

DeGroot, T., & Gooty, J. (2009). Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of Business and Psychology, 24*, 179–192. doi:10.1007/s10869-009-9098-0

DeGroot, T., & Kluemper, D. (2007). Evidence of predictive and incremental validity of personality factors, vocal attractiveness and the situational interview. *International Journal of Selection and Assessment, 15*, 30–39. doi:10.1111/j.1468-2389.2007.00365.x

Derous, E. (2007). Investigating personnel selection from a counseling perspective: Do applicants' and recruiters' perceptions correspond? *Journal of Employment Counseling, 44*, 60–72. doi:10.1002/j.2161-1920.2007.tb00025.x

Dipboye, R. L. (1982). Self-fulfilling prophecies in the selection recruitment interview. *Academy of Management Review, 7*, 579–587.

Dipboye, R. L. (1992). *Selection interviews: Process perspectives.* Cincinnati, OH: Southwestern.

Dipboye, R. L. (1994). Structured and unstructured interviews: Beyond the job-fit model. In G. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 12, pp. 79–123). Greenwich, CT: JAI Press.

Dipboye, R. L., Wooten, K., & Halverson, S. K. (2004). Behavioral and situational interviews. In J. C. Thomas (Ed.), *Comprehensive handbook of psychological assessment: Vol. 4. Industrial and organizational assessment* (pp. 297–316). Hoboken, NJ: Wiley.

Dougherty, T., Turban, D., & Callender, J. (1994). Confirming first impressions in the employment interview: A field study of interviewer behavior. *Journal of Applied Psychology, 79*, 659–665. doi:10.1037/0021-9010.79.5.659

Einhorn, L. J. (1981). Investigation of successful communicative behaviors. *Communication Education, 30*, 217–228. doi:10.1080/03634528109378473

Ellis, A. P. J., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type? *Journal of Applied Psychology, 87*, 1200–1208. doi:10.1037/0021-9010.87.6.1200

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–39315.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 231–254). New York, NY: Guilford Press.

Freund, T., Kruglanski, A. W., & Shpitzajzen, A. (1985). The freezing and unfreezing of impressional primacy: Effects of the need for structure and the fear of invalidity. *Personality and Social Psychology Bulletin, 11*, 479–487. doi:10.1177/0146167285114013

Friedman, H. S., & Riggio, R. E. (1981). Effect of individual differences in nonverbal expressiveness on transmission of emotion. *Journal of Nonverbal Behavior, 6*, 96–104. doi:10.1007/BF00987285

Furnham, A. (2008). HR professionals' beliefs about, and knowledge of, assessment techniques and psychometric tests. *International Journal of Assessment and Selection, 16*, 300–305. doi:10.1111/j.1468-2389.2008.00436.x

Graves, L. M. (1999). Gender bias in interviewers' evaluations of applicants: When and how does it occur? In G. N. Powell (Ed.), *Handbook of gender and work* (pp. 145–164). Thousand Oaks, CA: Sage.

Graves, L. M., & Powell, G. N. (1988). An investigation of sex discrimination in recruiters' evaluations of actual applicants. *Journal of Applied Psychology, 73*, 20–29. doi:10.1037/0021-9010.73.1.20

Harris, M. M., Dworkin, J. B., & Park, J. (1990). Preemployment screening procedures: How human resource managers perceive them. *Journal of Business and Psychology, 4*, 279–292. doi:10.1007/BF01125240

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1992). Primitive emotional contagion. In M. S. Clark (Ed.), *Review of personality and social psychology: Vol. 14. Emotion and social behavior* (pp. 151–177). Thousand Oaks, CA: Sage.

Hebl, M. R., Foster, J. B., Mannix, L. M., & Dovidio, J. F. (2002). Formal and interpersonal discrimination: A field study of bias toward homosexual applicants. *Personality and Social Psychology Bulletin, 28*, 815–825. doi:10.1177/0146167202289010

Herriot, P. (2002). Selection and self: Selection as a social process. *European Journal of Work and Organizational Psychology, 11*, 385–402. doi:10.1080/13594320244000256

Hollandsworth, J., Kazelskis, R., Stevens, J., & Dressel, M. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *Personnel Psychology, 32*, 359–367. doi:10.1111/j.1744-6570.1979.tb02140.x

Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology, 25*, 113–132. doi:10.1108/02683941011019339

Hosoda, M., Stone-Romero, E., & Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology, 56*, 431–462. doi:10.1111/j.1744-6570.2003.tb00157.x

Howard, D. J., & Gengler, C. (2001). Emotional contagion effects on product attitudes. *Journal of Consumer Research, 28*, 189–201. doi:10.1086/322897

Huffcutt, A., & Woehr, D. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior, 20*, 549–560. doi:10.1002/(SICI)1099-1379(199907)20:4<549::AID-JOB921>3.0.CO;2-Q

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184–190. doi:10.1037/0021-9010.79.2.184

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81*, 459–473. doi:10.1037/0021-9010.81.5.459

Janz, T., Hellervik, L., & Gilmore, D. C. (1986). *Behavior description interviewing: New, accurate, cost-effective*. Boston, MA: Allyn & Bacon.

Johnson, S. K. (2008). I second that emotion: Effects of emotional contagion and affect at work on leader and follower outcomes. *Leadership Quarterly, 19*, 1–19. doi:10.1016/j.leaqua.2007.12.001

Johnson, S. K. (2009). Do you feel what I feel? Mood contagion and leadership outcomes. *Leadership Quarterly, 20*, 814–827. doi:10.1016/j.leaqua.2009.06.012

Johnson, S. K., & Johnson, C. S. (2009). The secret life of mood: Causes and consequences of unconscious affect at work. In C. E. J. Härte, N. M. Ashkanasy, & W. J. Zerbe (Eds.), *Research on emotions in organizations* (pp. 103–121). Bingley, England: Emerald Group.

Johnson, S. K., Podratz, K. E., Dipboye, R. L., & Gibbons, E. (2010). Physical attractiveness biases in ratings of employment suitability: Tracking down the "beauty is beastly" effect. *Journal of Social Psychology, 150*, 301–318. doi:10.1080/00224540903365414

King, E., Madera, J., Hebl, M., Knight, J., & Mendoza, S. (2006). What's in a name? A multiracial investigation of the role of occupational stereotypes in selection decisions. *Journal of Applied Social Psychology, 36*, 1145–1159. doi:10.1111/j.0021-9029.2006.00035.x

King, E. B., Shapiro, J. R., Hebl, M. R., Singletary, S. L., & Turner, S. (2006). The stigma of obesity in customer service: A mechanism for remediation and bottom-line consequences of interpersonal discrimination. *Journal of Applied Psychology, 91*, 579–593. doi:10.1037/0021-9010.91.3.579

Kristof-Brown, A. (2000). Perceived applicant fit: Distinguishing between recruiters' perceptions of person–job and person–organization fit. *Personnel Psychology, 53*, 643–671. doi:10.1111/j.1744-6570.2000.tb00217.x

Kruglanski, A. W., & Mayseless, O. (1988). Contextual effects in hypothesis testing: The role of competing alternatives and epistemic motivations. *Social Cognition, 6*, 1–20.

Lance, A., & Wilson, S. (1997). Critical incidents technique. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 89–112). Palo Alto, CA: Davies-Black.

Larsen, R. J., & Kasimatis, M. (1990). Individual differences in entrainment of mood to the weekly calendar. *Journal of Personality and Social Psychology, 58*, 164–171. doi:10.1037/0022-3514.58.1.164

Latham, G., & Saari, L. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*, 569–573. doi:10.1037/0021-9010.69.4.569

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and validation of an interview faking behavior scale. *Journal of Applied Psychology, 92*, 1638–1656. doi:10.1037/0021-9010.92.6.1638

Levine, S. P., & Feldman, R. S. (2002). Women and men's nonverbal behavior and self-monitoring in a job interview setting. *Applied HRM Research, 7*, 1–14.

Macan, T., & Dipboye, R. (1990). The relationship of interviewers' preinterview impressions to selection and recruitment outcomes. *Personnel Psychology, 43*, 745–768. doi:10.1111/j.1744-6570.1990.tb00681.x

Marcus, B. (2009). "Faking" from the applicant's perspective: A theory of self-presentation in personnel selection settings. *International Journal of Selection and Assessment, 17*, 417–430. doi:10.1111/j.1468-2389.2009.00483.x

Markey, P. M., Funder, D. C., & Ozer, D. J. (2003). Complementarity of interpersonal behaviors in dyadic interactions. *Personality and Social Psychology Bulletin, 29*, 1082–1090. doi:10.1177/0146167203253474

Matarazzo, J. D., & Wiens, A. N. (1972). *The interview: Research on its anatomy and structure.* Chicago, IL: Aldine Atherton.

Maurer, S. D. (2002). A practitioner-based analysis of interviewer job expertise and scale format as contextual factors in situational interviews. *Personnel Psychology, 55*, 307–327. doi:10.1111/j.1744-6570.2002.tb00112.x

McCarthy, J. M., Van Iddekinge, C. H., & Campion, M. A. (2010). Are highly structured job interviews resistant to demographic similarity effects? *Personnel Psychology, 63*, 325–359. doi:10.1111/j.1744-6570.2010.01172.x

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616. doi:10.1037/0021-9010.79.4.599

Mook, D. (1983). In defense of external invalidity. *American Psychologist, 38*, 379–387. doi:10.1037/0003-066X.38.4.379

Morgeson, F. P., Reider, M. H., Campion, M. A., & Bull, R. A. (2008). Review of research on age discrimination in the employment interview. *Journal of Business and Psychology, 22*, 223–232. doi:10.1007/s10869-008-9066-0

Motowidlo, S. J., & Burnett, J. R. (1995). Aural and visual sources of validity in structured employment interviews. *Organizational Behavior and Human Decision Processes, 61*, 239–249. doi:10.1006/obhd.1995.1019

Neuberg, S. L. (1989). The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies. *Journal of Personality and Social Psychology, 56*, 374–386. doi:10.1037/0022-3514.56.3.374

Neumann, R., & Strack, F. (2000). "Mood contagion": The automatic transfer of mood between persons. *Journal of Personality and Social Psychology, 79*, 211–223. doi:10.1037/0022-3514.79.2.211

Newman, J., & Krzystofiak, F. (1979). Self-reports versus unobtrusive measures: Balancing method variance and ethical concerns in employment discrimination research. *Journal of Applied Psychology, 64*, 82–85. doi:10.1037/0021-9010.64.1.82

Nordstrom, C. R., Hall, R. J., & Bartels, L. K. (1998). First impressions versus good impressions: The effect of self-regulation on interview evaluations. *Journal of Psychology: Interdisciplinary and Applied, 132*, 477–491. doi:10.1080/00223989809599281

Pager, D., Western, B., & Bonikowski, B. (2009). Discrimination in a low-wage labor market: A field experiment. *American Sociological Review, 74*, 777–799. doi:10.1177/000312240907400505

Pager, D., Western, B., & Sugie, N. (2009). Sequencing disadvantage: Barriers to employment facing young Black and White men with criminal records. *Annals of the American Academy of Political and Social Science, 623*, 195–213. doi:10.1177/0002716208330793

Peeters, H., & Lievens, F. (2006). Verbal and nonverbal impression management tactics in behavior description and situational interviews. *International Journal of Selection and Assessment, 14*, 206–222. doi:10.1111/j.1468-2389.2006.00348.x

Polansky, N., Lippitt, R., & Redl, F. (1950). An investigation of behavioral contagion in groups. *Human Relations, 3*, 319–348. doi:10.1177/001872675000300401

Posthuma, R. A., Morgeson, F. P., & Campion, M. A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personnel Psychology, 55*, 1–81. doi:10.1111/j.1744-6570.2002.tb00103.x

Powell, R. S., & O'Neal, E. C. (1976). Communication feedback and duration as determinants of accuracy, confidence, and differentiation in interpersonal perception. *Journal of Personality and Social Psychology, 34*, 746–756. doi:10.1037/0022-3514.34.4.746

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48*, 289–308. doi:10.1111/j.1744-6570.1995.tb01758.x

Purkiss, L. S., Perrewé, P. L., Gillespie, T. L., Mayes, B. T., & Ferris, G. R. (2006). Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes, 101*, 152–167. doi:10.1016/j.obhdp.2006.06.005

Reilly, N., Bocketti, S., Maser, S., & Wennet, C. (2006). Benchmarks affect perceptions of prior disability in a structured interview. *Journal of Business and Psychology, 20*, 489–500. doi:10.1007/s10869-005-9005-2

Roberts, L., & Macan, T. (2006). Disability disclosure effects on employment interview ratings of applicants with nonvisible disabilities. *Rehabilitation Psychology, 51*, 239–246. doi:10.1037/0090-5550.51.3.239

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., & Bobko, P. (2002). Corrections for range restrictions in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87*, 369–376. doi:10.1037/0021-9010.87.2.369

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., & Schmit, M. (2005). Personality saturation in structured interviews. *International Journal of Selection and Assessment, 13*, 261–273. doi:10.1111/j.1468-2389.2005.00323.x

Rudolph, C., Wells, C., Weller, M., & Baltes, B. (2009). A meta-analysis of empirical studies of weight-based bias in the workplace. *Journal of Vocational Behavior, 74*, 1–10. doi:10.1016/j.jvb.2008.09.008

Russell, B., Perkins, J., & Grinnell, H. (2008). Interviewees' overuse of the word "like" and hesitations: Effects in simulated hiring decisions. *Psychological Reports, 102*, 111–118. doi:10.2466/pr0.102.1.111-118

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707–721. doi:10.1111/j.1744-6570.1997.tb00711.x

Sadler, P., Ethier, N., Gunn, G., Duong, D., & Woody, E. (2009). Are we on the same wavelength? Interpersonal complementarity as shared cyclical patterns during interactions. *Journal of Personality and Social Psychology, 97*, 1005–1020. doi:10.1037/a0016232

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11*, 299–324. doi:10.1080/13594320244000184

Schmidt, F., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274. doi:10.1037/0033-2909.124.2.262

Schmidt, F. L., & Zimmerman, R. D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89*, 553–561. doi:10.1037/0021-9010.89.3.553

Schreurs, B., Derous, E., De Witte, K., Proost, K., Andriessen, M., & Glabeke, K. (2005). Attracting potential applicants to them: The effects of initial face-to-face contacts. *Human Performance, 18*, 105–122. doi:10.1207/s15327043hup1802_1

Sczesny, S., & Stahlberg, D. (2002). The influence of gender-stereotyped perfumes on leadership attribution. *European Journal of Social Psychology, 32*, 815–828. doi:10.1002/ejsp.123

Sherman, J. W., & Frost, L. A. (2000). On the encoding of stereotype-relevant information under cognitive load. *Personality and Social Psychology Bulletin, 26*, 26–34. doi:10.1177/0146167200261003

Silvester, J., & Anderson, N. (2003). Technology and discourse: A comparison of face-to-face and telephone employment interviews. *International Journal of Selection and Assessment, 11*, 206–214. doi:10.1111/1468-2389.00244

Singer, M., & Bruhns, C. (1991). Relative effect of applicant work experience and academic qualification on selection interview decisions: A study of between-sample generalizability. *Journal of Applied Psychology, 76*, 550–559. doi:10.1037/0021-9010.76.4.550

Snyder, M., & Klein, O. (2005). Construing and constructing others: On the reality and the generality of the behavioral confirmation scenario. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems, 6*, 53–67. doi:10.1075/is.6.1.05sny

Stevens, C., & Kristof, A. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology, 80*, 587–606. doi:10.1037/0021-9010.80.5.587

Stevens, C. K. (1998). Antecedents of interview interactions, interviewers' ratings, and applicants' reactions. *Personnel Psychology, 51*, 55–85. doi:10.1111/j.1744-6570.1998.tb00716.x

Stewart, G. L., Dustin, S. L., Barrick, M. R., & Darnold, T. C. (2008). Exploring the handshake in employment interviews. *Journal of Applied Psychology, 93*, 1139–1146. doi:10.1037/0021-9010.93.5.1139

Straus, S. G., Miles, J. A., & Levesque, L. L. (2001). The effects of videoconference, telephone, and face-to-face media on interviewer and applicant judgments in employment interviews. *Journal of Management, 27*, 363–381. doi:10.1016/S0149-2063(01)00096-4

Sullins, E. S. (1991). Emotional contagion revisited: Effects of social comparison and expressive style on mood convergence. *Personality and Social Psychology Bulletin, 17*, 166–174. doi:10.1177/014616729101700208

Swann, W. B., & Ely, R. J. (1984). A battle of wills: Self-verification versus behavioral confirmation. *Journal of Personality and Social Psychology, 46*, 1287–1302. doi:10.1037/0022-3514.46.6.1287

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75*, 277–294. doi:10.1348/096317902320369712

Tengler, C., & Jablin, F. (1983). Effects of question type, orientation, and sequencing in the employment screening interview. *Communication Monographs, 50*, 245–263. doi:10.1080/03637758309390167

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology, 43*, 22–34. doi:10.1037/0022-3514.43.1.22

Tullar, W. L. (1989). Relational control in the employment interview. *Journal of Applied Psychology, 74*, 971–977. doi:10.1037/0021-9010.74.6.971

Van Iddekinge, C. H., McFarland, L. A., & Raymark, P. H. (2007). Antecedents of impression management use and effectiveness in a structured interview. *Journal of Management, 33*, 752–773. doi:10.1177/0149206307305563

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured

employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology, 90*, 536–552. doi:10.1037/0021-9010.90.3.536

Van Iddekinge, C. H., Sager, C. E., Burnfield, J. L., & Heffner, T. S. (2006). The variability of criterion-related validity estimates among interviewers and interview panels. *International Journal of Selection and Assessment, 14*, 193–205. doi:10.1111/j.1468-2389.2006.00352.x

Wade, K., & Kinicki, A. (1997). Subjective applicant qualifications and interpersonal attraction as mediators within a process model of interview selection decisions. *Journal of Vocational Behavior, 50*, 23–40. doi:10.1006/jvbe.1996.1538

Walters, L. C., Miller, M. R., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *International Journal of Aviation Psychology, 3*, 25–38. doi:10.1207/s15327108ijap0301_2

Whetzel, D., Baranowski, L., Petro, J., Curtin, P., & Fisher, J. (2003). A written structured interview by any other name is still a selection instrument. *Applied H. R. M. Research, 8*, 1–16.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275–290. doi:10.1111/j.2044-8325.1988.tb00467.x

Williamson, L., Campion, J., Malos, S., Roehling, M., & Campion, M. (1997). Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology, 82*, 900–912. doi:10.1037/0021-9010.82.6.900

Wish, M., Deutsch, M., & Kaplan, S. (1976). Perceived dimensions of interpersonal relations. *Journal of Personality and Social Psychology, 33*, 409–420. doi:10.1037/0022-3514.33.4.409

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology, 10*, 109–120. doi:10.1016/0022-1031(74)90059-6

Ziegert, J. C., & Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology, 90*, 553–562. doi:10.1037/0021-9010.90.3.553

CHAPTER 28

PERSONALITY MEASUREMENT AND USE IN INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY

Leaetta M. Hough and Brian S. Connelly

Personality variables are important determinants of behavior in virtually all aspects of life. Yet, in its quest to understand behavior in the workplace, industrial and organizational (I/O) psychology has a history of being highly critical of personality variables and their measurement. Even today, in the 21st century, controversy and unease about how personality influences work behavior exist.

It has been a tumultuous and contentious history. Every 10 or so years, the same issues reappear, albeit wrapped in somewhat different clothes. One recurring criticism is that personality variables account for only a small amount of variability in work performance. This criticism leads to, for all practical purposes, excluding personality variables from models of the determinants of work behavior and performance. A second recurring criticism is that even if personality variables are determinants of work behavior, they cannot be measured in settings in which the scores are used to make decisions about people because respondents (e.g., job applicants) will intentionally distort their responses, rendering the scores essentially useless. These two criticisms have plagued, and continue to plague, personality variables and their use throughout the history of I/O psychology.

This chapter addresses these issues, summarizing what is known about the role of personality variables in understanding behavior and how to measure them to ensure their usefulness in the workplace. When appropriately analyzed and measured, personality variables are important determinants of work performance.

WHY PERSONALITY IS IMPORTANT TO INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY

Personality variables contribute significantly to the understanding of work behavior. They contribute to models of the determinants of work performance and to greater accuracy in predicting work behavior and performance. They are important determinants of overall job performance as well as more specific components of work behavior and work life. Moreover, when incorporated into personnel selection systems, they produce a workforce that better matches the ethnic and gender demographics of the community. Without question, personality variables are important for both the science and the practice of I/O psychology.

Better Understanding of Work Behavior

I/O psychology is the science of work and organizational behavior; empirical studies have guided the development and revision of its models and theories. Over the years, increasingly more complex models of the determinants of work performance and work adjustment have been postulated and researched. Today, considerable evidence exists in the form of validity regarding the components of the work performance and work adjustment models.

Industrial and organizational psychology models of work performance and work adjustment. Early theorists hypothesized that Performance = Ability × Motivation (see J. P. Campbell & Pritchard, 1976). Personality was not part of the equation, nor was

DOI: 10.1037/14047-028
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

it considered part of motivation. The paradigm in academic circles in the 1960s, 1970s, and much of the 1980s excluded personality variables as playing a role in determining work behavior. It was, as Hough and Ones (2001) described, a dark age for personality variables in I/O psychology and for most of psychology. During those dark ages, models of the determinants of work performance became more specific and included cognitive ability, job knowledge, and task proficiency (e.g., Hunter, 1983) but not personality. Job experience was shortly added (e.g., Schmidt, Hunter, & Outerbridge, 1986). Not until the 1990s, after construct-oriented reviews of personality–criterion relationships advanced understanding (Barrick & Mount, 1991; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990), was a personality variable included in a model of the determinants of work performance (e.g., Borman, White, Pulakos, & Oppler, 1991). The Borman et al. (1991) model incorporated two facets of Conscientiousness—dependability and achievement orientation—and accounted for more than twice the variance in job performance ratings than the Hunter (1983) model. Schmidt and Hunter (1992) summarized evidence for a similar model that included general mental ability, job experience, and conscientiousness as determinants of job performance, concluding that (a) general mental ability is a significant causal determinant of the acquisition of job knowledge and an indirect determinant of job performance through its influence on job knowledge acquisition and (b) Conscientiousness is a direct causal determinant of job performance and an indirect determinant of job performance through its effect on job knowledge acquisition.

During the 1990s, many I/O psychologists conceded that Conscientiousness might be a useful predictor of job performance across work settings, but the importance of any other personality variable was suspect until more complex thinking about work performance expanded the criterion space. In addition to overall job performance, work performance constructs such as task performance and contextual performance (also known as *organizational citizenship*) were recognized as important components of work performance (e.g., Motowidlo, Borman, & Schmit, 1997; Organ & Ryan, 1995). Expanding the criterion space to include contextual performance helped solidify the importance of Conscientiousness. With a more nuanced understanding of the criterion space, personality variables other than Conscientiousness also emerged as important determinants of work performance. Even the maligned personality variable Openness to Experience gained stature as an important variable in understanding innovation and creativity (Bartram, 2005; Hough, 1992; Hough & Dilchert, 2007).

More complex thinking about the nature of the relationships between variables resulted in the inclusion of additional personality variables as well as their role as moderator and mediator variables. Oswald and Hough (2011) provided a generic process model that integrates goal setting, motivation, goal orientation, episodic performance behavior, and revisions of goals and episodic performance behavior over time as well as moderator variables such as national, organizational, and team culture; equipment; measurement method; validation strategy; and rater perspective. Their model elucidates some of the factors and processes in the "black box" of personality–performance validities. Johnson and Hezlett (2008) proposed a more specific content model to explain the psychological processes that influence work performance in organizations. Their model differentiates between distal variables (such as personality, ability, experience, and organizational context) and proximal variables (such as self-efficacy, goal setting, autonomy, and stress) of work performance. Other process models have emerged as well to explain, for example, training motivation (e.g., Colquitt, LePine, & Noe, 2000). These more detailed models deserve attention as well as rigorous research to bootstrap I/O psychology's way to a better and more sophisticated understanding of the determinants of work performance and adjustment.

Validity. One of the important advances in I/O psychology has been a greater focus on constructs in both predictor and criterion domains. The emphasis on constructs has enabled researchers to summarize criterion-related validities according to predictor–criterion construct combinations, such as Conscientiousness–contextual performance, producing theoretically meaningful summaries of the relationships between personality constructs

and criterion constructs. Without question, meta-analytic summaries of the relationship between personality constructs and work-related constructs produced insights into the importance of personality in the workplace.

One of the most important insights—although some might argue it is not an insight but obvious—is this: Personality constructs relate to work performance and work adjustment constructs differently. On the basis of dozens of meta-analyses (many of which are referenced later), one conclusion is very clear: Personality variables influence work behavior outcomes. In contrast to the cognitive ability domain, in which a meaningful general mental ability variable (*g*) exists and it and its subcomponents correlate similarly with criteria, no general personality variable subsumes all lower level personality variables and correlates usefully with a variety of criteria. These different personality–criterion construct relationships are discussed in more detail later. For now, the many areas in which personality variables influence work behavior include the following:

- occupational interests, occupational and career choice, career indecision, and career aspirations (e.g., Dawis & Lofquist, 1984; De Fruyt & Mervielde, 1997; Holland, 1997; O'Brien & Fassinger, 1993; Rainey & Borders, 1997; P. L. Schneider, Ryan, Tracey, & Rounds, 1996; Spector, Jex, & Chen, 1995);
- motivation to learn (e.g., Colquitt et al., 2000);
- e-learning (e.g., Orvis, Brusso, Wasserman, & Fisher, 2010);
- training outcomes (e.g., Barrick & Mount, 1991; Colquitt et al., 2000; Hough, 1992; Hough et al., 1990; Hurtz & Donovan, 2000; Ones, Viswesvaran, & Reiss, 1996; Salgado, 1997);
- educational outcomes (e.g., Connelly & Ones, 2010; Goldberg, Sweeney, Merenda, & Hughes, 1998; Hough, 1992; Hough et al., 1990; Lievens, Ones, & Dilchert, 2009; Paunonen, 2003; Poropat, 2009);
- job knowledge (e.g., Borman et al., 1991);
- employment status and job search activities (e.g., DeFruyt & Mervielde, 1999; Wanberg, Hough, & Song, 2002; Wanberg, Watt, & Rumsey, 1996);
- organizational choice (e.g., Jordan, Herriot, & Chalmers, 1991; B. Schneider, Smith, Taylor, & Fleenor, 1998);
- overall job performance (e.g., Barrick & Mount, 1991; Bartram, 2005; Connelly & Ones, 2010; Dudley, Orvis, Lebiecki, & Cortina, 2006; Hogan & Holland, 2003; Hough, 1992; Hough et al., 1990; Hurtz & Donovan, 2000; Judge & Bono, 2001; Ones, Dilchert, Viswesvaran, & Judge, 2007; Ones, Viswesvaran, & Schmidt, 1993; Ones & Viswesvaran, 2001; Salgado, 1997; Tett, Jackson, Rothstein, & Reddon, 1994);
- job and career satisfaction (e.g., Judge & Bono, 2001; Judge, Heller, & Mount, 2002; Moorman & Podsakoff, 1992; Ng, Eby, Sorensen, & Feldman, 2005; Thoresen, Kaplan, Barsky, de Chermont, & Warren, 2003);
- occupational and career attainment, level, advancement, success, and salary (e.g., Judge, Higgins, Thoresen, & Barrick, 1999; Ng et al., 2005; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007);
- entrepreneurship (e.g., Rauch & Frese, 2007; Zhao & Seibert, 2006);
- leadership, leadership emergence, and transformational leadership (e.g., Bartram, 2005; Bono & Judge, 2004; Derue, Nahrgang, Wellman, & Humphrey, 2011; Judge, Bono, Ilies, & Gerhardt, 2002);
- managerial effectiveness, promotion, level, and salary (e.g., Barrick & Mount, 1991; Dilchert & Ones, 2008; Dudley et al., 2006; Hough, Ones, & Viswesvaran, 1998);
- sales effectiveness (e.g., Barrick & Mount, 1991; Dudley et al., 2006; Hough, 1992; McCune et al., 2007; Vinchur, Schippmann, Switzer, & Roth, 1998);
- customer service (e.g., Dudley et al., 2006; Frei & McDaniel, 1998; Hogan, Hogan, & Busch, 1984; Hurtz & Donovan, 2000; Ones & Viswesvaran, 2001);
- interpersonal effectiveness (e.g., Mount, Barrick, & Stewart, 1998; Robertson & Kinder, 1993; cf. R. J. Schneider, Ackerman, & Kanfer, 1996);
- skilled and semiskilled worker job performance (e.g., Barrick & Mount, 1991; Dudley et al., 2006);

- expatriate job performance (e.g., Mol, Born, Willemsen, & Van Der Molen, 2005);
- combat effectiveness (e.g., Hough, 1992);
- goal setting and not procrastinating (e.g., Judge & Ilies, 2002; Steel, 2007);
- innovation and creativity (e.g., Bartram, 2005; Feist, 1998; Hough, 1992; Hough & Dilchert, 2007; Robertson & Kinder, 1993);
- contextual performance such as organizational citizenship, dedication, interpersonal facilitation, and altruism (e.g., Berry, Ones, & Sackett, 2007; Borman, Penner, Allen, & Motowidlo, 2001; Dudley et al., 2006; Hough, 1992; Hurtz & Donovan, 2000; LePine, Erez, & Johnson, 2002; Organ & Ryan, 1995);
- counterproductive work behavior, including theft, property damage, lateness, absenteeism, disciplinary problems, substance abuse, and violence on the job (e.g., Berry et al., 2007; Dudley et al., 2006; Hough, 1992; Hough et al., 1990; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Ones & Viswesvaran, 1998b; Ones et al., 1993; Ones, Viswesvaran, & Schmidt, 2003; Schmidt, Viswesvaran, & Ones, 1997);
- tenure and turnover (e.g., Barrick & Mount, 1991; Thoresen et al., 2003; Zimmerman, 2008);
- workplace safety and accidents (e.g., J. Arthur, Barrett, & Alexander, 1991; Christian, Bradley, Wallace, & Burke, 2009; Hansen, 1988, 1989; Ones & Viswesvaran, 1998a); and
- teamwork, team cohesion, and team performance (e.g., Hogan & Holland, 2003; Hough, 1992; Neuman & Wright, 1999; Peeters, Van Tuijl, Rutte, & Reymen, 2006).

The level of validity for an individual personality variable predicting a particular criterion is typically moderate in size. Nonetheless, these levels are highly valuable in practice especially when they are combined with other personality variables that are also theoretically relevant to the criterion. Hough and Furnham (2003) summarized the results of many meta-analyses and noted that (a) although Conscientiousness correlates most highly with overall performance for many jobs, other personality variables often correlate more highly with specific criterion constructs; (b) the level of validity for personality variables varies depending on the criterion construct; (c) even within a personality–criterion construct combination, meta-analytic validities vary, likely as a result of the type of job and setting; and (d) compound (complex) personality variables often correlate highest with complex criteria, criteria that are similarly complex and theoretically relevant to the predictor. Meta-analyses reported since Hough and Furnham, many of which were cited in the preceding list, have reinforced Hough and Furnham's conclusions.

Personality variables clearly help psychologists understand behavior and performance and warrant inclusion in models of work performance and adjustment. Personality variables typically increment the level of overall prediction when combined with variables such as cognitive ability variables that do not correlate with personality variables, and, as discussed in the next section, they do so with less collateral damage.

Less Adverse Impact on Protected Groups for Employment Decisions

Adverse impact is an important concept for I/O psychologists involved in personnel selection. Adverse impact is calculated by comparing the selection ratio of one group to another group. (A selection ratio is calculated by dividing the number of, e.g., African Americans hired by the total number of African American applicants.) If the ratios are not equal, it means an organization's hiring decisions affect groups of applicants differently. One consequence of differential hiring rates is an organizational workforce that does not reflect the makeup of the community at large.

In the United States, if the selection ratio of a protected group is less than 80% of the selection ratio of another group, adverse impact is considered to be present. Enforcement agencies in the United States use the 80% rule of thumb to evaluate whether employment decision-making processes and tools, such as an employment test, advantage one group over another group.

When adverse impact exists, organizations are at risk in a variety of ways. In the United States, where several laws protect minorities, women, and older people against employment practices that produce

adverse impact but lack job relatedness (validity), close scrutiny of an organization's employment hiring processes and decisions is likely and legal charges for discriminatory hiring practices are possible. Even if legal scrutiny and action were not possible consequences, the nonrepresentative nature of the organization's workforce can easily render it less competitive from marketing and sales perspectives and in candidate recruitment.

An important determinant of adverse impact is a mean-score difference between groups on a decision-making tool. As a result, organizations and personnel psychologists are keenly interested in mean-score differences between Whites and minorities, men and women, and younger and older people on measures of individual characteristics, especially when those measures are used to hire people.

Two large-scale meta-analyses of mean-score differences on personality variables between Whites and various ethnic groups produced similar results: small, if any, differences (Foldes, Duehr, & Ones, 2008; Hough, Oswald, & Ployhart, 2001). Both studies reported mean *d*, which is the effect size (standardized mean-score difference) of the mean-score difference. At the broad level of personality measurement (e.g., Emotional Stability, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness), for those comparisons in which the minority groups (Blacks and Hispanics) had a sample size of 1,500 or more, differences ranged from approximately 0.00 to 0.20. Absolute average *d* across these five personality constructs and the two meta-analyses was 0.06, with Blacks and Hispanics scoring higher than Whites about one third of the time. (In contrast, on measures of general mental ability, average *d* is approximately 1.0 for White–Black comparisons and 0.5 for the White–Hispanic comparisons, with Whites scoring higher on average than Blacks and Hispanics; Arvey et al., 1994; Hough et al., 2001.) Hough et al. (2001) and Foldes et al. (2008) found a few small differences between ethnic groups at the facet level and, sometimes, in opposite directions, even though the facets are considered by many to be part of the same larger personality construct. For example, although Blacks on average score about one third of a standard deviation lower than Whites on sociability, they score about the same (Foldes et al., 2008) or slightly higher ($d = 0.12$; Hough et al., 2001) than Whites on dominance, even though both facets are subsumed under Extraversion.

Gender comparisons present a somewhat different picture. One large-scale meta-analysis (Hough et al., 2001) compared men's and women's mean scores on broad-level and facet-level personality variables. At the broad level of personality measurement (e.g., Emotional Stability, Extraversion, Openness to Experience, Agreeableness, Conscientiousness, and Integrity) with sample sizes averaging about 130,000 for women and 140,000 for men, women scored 0.39 standard deviation ($d = 0.39$) higher than men on Agreeableness and 0.24 standard deviation ($d = 0.24$) lower than men on adjustment, lending support to some common stereotypes of the differences between men and women. Interesting differences were also found at the facet level; for two facets of Extraversion, that is, sociability and dominance, women on average scored about 0.25 standard deviation lower than men on dominance but about 0.10 standard deviation higher than men on sociability. An examination of two facets of Conscientiousness, that is, dependability and achievement, produced interesting results as well. Women scored about 0.5 standard deviation higher than men on dependability but essentially the same as men on achievement.

Age comparisons also present an interesting picture. One large-scale meta-analysis (Hough et al., 2001) compared mean scores on broad- and facet-level personality variables for working adults age 40 years or younger with those for working adults older than age 40. They found a small difference ($d = 0.21$) on Agreeableness, with older working adults scoring higher on Agreeableness. At the facet level, older adults scored higher ($d = 0.49$) on dependability than did younger working adults.

These broad- and facet-level differences are noteworthy. For example, Agreeableness predicts teamwork, customer service, and interpersonal facilitation. If selection was based only on Agreeableness, minority and White hiring rates would likely be similar, but hiring rates for men and women and older and younger applicants would likely be different. More women than men would likely be hired, and a slightly larger number of older than younger

people would likely be hired. Similarly, if hiring were based on dependability, which predicts overall job performance, dedication, organizational citizenship, and lack of counterproductive work behavior, again somewhat more women and older applicants would be hired. These findings are not likely to lead to legal ramifications in the United States because men and younger people are not protected groups (women and older applicants are). Measures of dominance (a facet of Extraversion) would, however, lead to a somewhat different outcome for men and women; somewhat more men than women would be hired.

Personality variables that are included in personnel selection systems should be carefully selected or developed to correspond to actual work requirements and relevant work criteria. Measuring a personality variable at an inappropriate level can contribute to or ameliorate adverse impact and, as described next, contribute to (or reduce) predictive validity.

WHAT TO MEASURE: RELEVANT PERSONALITY CONSTRUCTS FOR INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY

Traits have provided fundamental building blocks enabling significant advances in I/O psychology's quest to understand the determinants of work outcomes. Much of the field's increased knowledge flows from a trait paradigm, a paradigm that has relied on at least two important requirements: (a) a taxonomic structure of personality and (b) consistency of behavior across situations. These two requirements allow aggregation and generalization of knowledge.

At least three basic approaches have been used to identify personality structures—theoretical, lexical, and nomological-web clustering. Theoretical approaches are guided by theory. Freud's id and superego are examples of well-known constructs that are the product of a theory-based approach. I/O psychologists have not, in general, incorporated theory-based constructs into their models of the determinants of work outcomes. They have, however, incorporated lexical-based traits into their models. The five-factor model, which has dominated personality research in I/O psychology and elsewhere, is based on the lexical approach. A third, relatively little known nomological-web clustering approach is an alternative strategy for identifying constructs relevant to I/O psychology.

Lexical Approach and the Five- and Six-Factor Models

Many psychologists embrace the five factors of the five-factor model of personality (Goldberg, 1993; Tupes & Christal, 1961/1992) as the basic, universal set of variables that describe personality. Its research paradigm is a lexical (language) approach to identifying the basic units of personality. Galton (1884), who was the first to identify and catalog personality descriptors, postulated that personality traits can be captured in the words people use to describe each other. Others, such as Allport and Odbert (1936), Cattell (1943), and Fiske (1949), continued this line of research, with Tupes and Christal (1961/1992) identifying a five-factor model that is highly similar to today's five-factor model—Emotional Stability (Neuroticism), Extraversion, Openness to Experience, Conscientiousness, and Agreeableness. (See Hough & Schneider, 1996, for a history of the five-factor model.)

The five-factor model has been refined over the years as dozens of studies have examined its stability across raters, ethnic groups, gender, cultures, languages, time, and type of factor extraction and rotation method. Although many studies have supported a conclusion that the five-factor model is a robust, universal model of the structure of personality, other studies have not supported that conclusion, especially for non-English languages. (See Oswald & Hough, 2011, for a discussion of conflicting results.)

A sixth factor. In a reanalysis of the data from studies examining the factor structure of lexical terms across seven languages (Dutch, French, German, Hungarian, Italian, Korean, Polish), Ashton et al. (2004) found a sixth factor—Honesty–Humility—in addition to analogs of each of the Big Five factors. They named their model the HEXACO model. The six-factor solution not only better accounts for the variance in lexical ratings

across languages, it better accounts for variance in predictor–criterion relationships. Across four cross-cultural samples, the six factors accounted for approximately 10% to 15% more variance in the prediction of workplace delinquency than the five-factor model (Lee, Ashton, & deVries, 2005). Table 28.1 defines the five and six factors, highlighting the similarities and differences between the two models.

Many of the meta-analyses referenced in the Validity section earlier in this chapter aggregated criterion-related validities according to the five-factor model. Analyzing data using the five-factor model of personality variables allowed important, hitherto obscured, relationships between personality constructs and work-related criterion constructs to emerge. Yet, other taxonomic structures such as the HEXACO model appear to better account for the relationships among personality variables and between personality–criterion construct combinations.

Facet-level measurement. Hough and her colleagues (Hough, 1989, 1992, 1998; Hough & Dilchert, 2010; Hough & Oswald, 2000, 2005, 2008; Hough & Schneider, 1996; Oswald & Hough, 2011; R. J. Schneider, Hough, & Dunnette, 1996) have argued that the factors of the five-factor model and the six factors of the HEXACO model are often too heterogeneous and that the facets that make up these broad factors differ in their relationships with important work-related constructs, differences that the broader factors mask. More and more research has demonstrated the accuracy of this observation (e.g., Ashton, 1998; Dudley et al., 2006; Kwong & Cheung, 2003; LePine, Colquitt, & Erez, 2000; Moon, Hollenbeck, Marinova, & Humphrey, 2008; Paunonen & Nicol, 2001; Roberts, Chernyshenko, Stark, & Goldberg, 2005; Stewart, 1999; Vinchur et al., 1998; Warr, Bartram, & Martin, 2005).

Hough and her colleagues (Hough, 1989, 1992, 1998; Hough & Dilchert, 2010; Hough & Oswald, 2000, 2005, 2008; Hough & Schneider, 1996; Oswald & Hough, 2011; R. J. Schneider, Hough, & Dunnette, 1996) also bolstered their argument with evidence that (a) subgroup (e.g., ethnic, gender, and age groups) differences in mean scores exist for some facets but not other facets within the same factor or (b) one subgroup scores significantly higher than

TABLE 28.1

Description of Five-Factor and the Six-Factor (HEXACO) Models of Personality

HEXACO model (6 factors)		Five-factor model	
Factor	**Definition[a]**	**Factor**	**Definition[b]**
H: Honesty—Humility	Sincere, modest, nongreedy, fair minded		
E: Emotional Stability (reversed)–Emotionality	Anxious, fearful, dependent, sentimental	Neuroticism	Anxious, angry, depressed, self-conscious, immoderate, vulnerable
X: Extraversion–Surgency	Expressive, socially bold, social, lively	Extraversion	Gregarious, friendly, assertive, active, cheerful, excitement seeking
A: Agreeableness	Forgiving, gentle, flexible, patient	Agreeableness	Trustworthy, moral, altruistic, cooperative, modest, sympathetic
C: Conscientiousness	Organized, diligent, perfectionist, prudent	Conscientiousness	Orderly, dutiful, achievement striving, self-disciplined, cautious
O: Openness to Experience	Creative, curious, unconventional, aesthetic appreciation	Openness to Experience	Imaginative, artistic interests, intellectual, liberal, adventurous, emotional

[a]From Lee and Ashton (2004). [b]From International Personality Item Pool (Goldberg, 1999).

another subgroup on one facet but significantly lower on another facet of the same factor. Examples of such differences were reported in the Less Adverse Impact on Protected Groups for Employment Decisions section: African Americans scored lower than Whites on sociability but higher than Whites on dominance, both facets of Extraversion. Similarly, subgroups scored differently on dependability and achievement, two facets of Conscientiousness: Older working adults scored higher than younger working adults on dependability but lower on achievement.

In short, measurement at the more narrow level often reveals more useful information than measurement at the factor level. Moreover, scores at the facet level can be aggregated to form more heterogeneous constructs at the five- and six-factor levels. Even higher level, more heterogeneous or compound variables such as managerial potential or social service orientation can be formed from facet-level scales.

Nomological-Web Clustering Approach

The lexical approach to discovering the universal taxonomic structure of personality has provided I/O psychology with very useful constructs. There can be no doubt that the five-factor model of personality has improved models of the determinants of work outcomes. Nonetheless, there are problems with the five-factor and HEXACO (six-factor) models. The nomological-web clustering approach is an alternative method of developing a useful set of constructs for I/O psychology. The constructs identified with the nomological-web clustering approach better account for the relationships among personality variables and personality–criterion construct combinations than either the five-factor or the HEXACO models. Instead of relying solely on ratings based on personality language descriptions to form constructs, the nomological-web clustering approach examines the pattern of relationships of a variable with other personality variables as well as external variables such as criterion constructs to form clusters based on the similarity of the nomological nets of the target variables. Its focus on both personality–personality relationships and personality–criterion relationships differentiates it from the lexical approach.

Hough and Ones (2001) developed a set of constructs using this approach and invited others to refine it. See Table 28.2 for a list of their constructs. Since then, at least two studies (Dudley et al., 2006; Foldes et al., 2008) have used the Hough and Ones taxonomy to contribute to a greater understanding of personality variables and their role in I/O psychology.

MEASUREMENT METHODS: SELF-REPORT

Personality characteristics in the workplace are typically measured with self-report, in which the respondent endorses or chooses adjectives or statements as being self-descriptive. Although self-assessments are known to be flawed (Mabe & West, 1982), data presented earlier in the Validity section have indicated that self-report measures correlate with important life and workplace outcomes. After a careful review of the literature, Chan (2009) questioned the negative conclusions regarding self-report measures, concluding that the evidence against the usefulness of self-reports in organizational and social sciences is more urban legend than reality. Nonetheless, in high-stakes testing situations, such as personnel selection, in which applicants are understandably motivated to present themselves in a desirable way, many researchers and employers still consider self-reports suspect and express concern about the effect of intentional distortion on the usefulness of self-report measures.

Potential Threat to Validity: Intentional Distortion

Concerns that applicants might intentionally describe themselves in an overly positive way on self-report measures are as old as personality measurement. Certainly, the lure of a job offer represents an especially strong incentive to misrepresent oneself. The term *high-stakes testing,* often used to describe real-life personnel selection situations, captures the importance of the outcome for individuals and organizations alike. In spite of meta-analytic evidence that provides valuable information, researchers have widely differing views about the severity and prevalence of faking, its psychometric effects among applicant samples, and its effects on validity.

In both I/O and personality psychology, the research literature on faking is voluminous and

TABLE 28.2

Constructs Generated via Nomological Web Clustering

Big Five and facets	Compound variables	Combination of these Big Five factors
Emotional Stability (ES)	Optimism	+ES +EX
Self-esteem	Intraception	+ES +OE
Low anxiety	Trust	+ES +A
Even tempered	Self-control	+ES +C
Extraversion (EX)	Reflective	−EX +OE
Dominance	Modesty	−EX +A
Sociability	Warmth	+EX +A
Activity–energy level	Ambition	+EX +C
Openness to Experience (OE)	Autonomy	+EX −C
Complexity	Tolerance	+OE +A
Culture–artistic	Traditionalism	−OE +C
Creativity–innovation	Lack of aggression	+A +C
Change–variety	Fair and stable leadership	+ES +EX +C
Curiosity–breadth	Self-destructive autonomy	−ES +EX −C
Intellect	Socialization	+ES +A +C
Agreeableness (A)	Thrill Seeking	+EX +OE −C
Nurturance	Democratic	+EX +A +C
Conscientiousness (C)	Achievement via independence	+ES +EX +OE +C
Achievement		
Dependability		
Moralistic		
Order		
Persistence		
Cautiousness–impulse control vs. risk taking–impulsive	Masculinity (rugged individualism)	

Note. Constructs from Hough and Ones (2001).

growing, with a slew of methodologies groping to understand and identify job applicants who have distorted their self-descriptions. One can easily find empirical evidence to support simplistic claims such as "faking doesn't matter" as well as "faking renders personality tests useless." The data and results are more complex; broad, sweeping conclusions about faking are in general unwarranted.

Research designs, measurement methods, and administration instructions make a difference in the results obtained; each affects the amount of distortion in self-reports, the veracity of self-descriptions, and the consequences of distortion. Five basic questions need to be answered:

1. Can respondents fake?
2. Can faking be detected?
3. What are the consequences of faking?
4. Can faking be deterred?
5. Do job applicants fake?

Can respondents fake? A wealth of research has compared responses of individuals instructed to fake good (and fake bad) with responses of individuals responding honestly. Although such directed-faking designs are criticized for their lack of applicability to actual applicant settings, these designs afford insight into respondents' ability to fake and the effects of doing so. Here, there is no debate: Respondents can fake self-report inventories in both positive and negative directions (Stanush, 1997; Viswesvaran & Ones, 1999). Directed-faking studies have revealed that psychometric properties of tests

change substantially when respondents intentionally distort their responses:

- Mean scores change significantly from honest to fake conditions: Mean scores in fake-good conditions generally increase from about 0.5 to a full standard deviation, whereas in fake-bad conditions mean scores change even more but in the opposite direction. The differences are even more extreme when results from within-subject study designs are compared with between-subjects study designs, with differences between honest and faking conditions greater for within-subject designs than for between-subject designs (Hooper, 2007; Stanush, 1997; Viswesvaran & Ones, 1999).
- Faking creates a ceiling effect and reduces variability substantially, with standard deviations decreasing by about 75% in directed-faking conditions (Hooper, 2007).
- Correlations between honest and faked responses are generally weak, indicating that faking changes the rank ordering of individuals (Ellingson, Sackett, & Hough, 1999). Studies have revealed that, even in directed-faking conditions, respondents fake to different degrees. Individual differences in faking introduce construct-irrelevant variance in test scores that results in decreased construct validity.
- Faking on multidimensional personality inventories collapses the factor structure of inventories, often to a single factor, in directed-faking conditions (Ellingson, Smith, & Sackett, 2001; Zickar & Robie, 1999; Schmit & Ryan, 1993). Previously independent factors merge, forming one large factor.

In short, respondents can fake their responses to self-report questions when instructed to do so. Directed-faking studies have indicated that the psychometric consequences are significant, and they are negative. Although directed-faking studies can provide a basis for understanding intentional distortion, they do not necessarily provide evidence about the behavior of job applicants in real-life selection situations.

Can faking be detected? Personality researchers have long been interested in identifying those who distort their responses and have explored a variety of approaches to measuring socially desirable responding. Traditional personality inventories have embedded unlikely-virtues items (e.g., "I have never told a lie"), with the assumption that respondents endorsing many of these items are distorting their responses. Classic social desirability measures became more nuanced when Paulhus (1984) distinguished effortful attempts at social desirability (impression management) from unconsciously held, overly positive self-views (self-deceptive enhancement).

Two forms of evidence from directed-faking studies have suggested that unlikely-virtues scales detect intentional distortion. First, in directed-faking studies, the meta-analytically derived correlation between typical unlikely-virtues scales and other personality scales is .09 in honest conditions but .34 in faking conditions (Stanush, 1997). Second, the difference in mean scores for typical unlikely-virtues scales in fake-good conditions compared with honest conditions are significantly higher than the difference in mean scores for other scales in fake-good conditions compared with honest conditions, suggesting that typical unlikely-virtues scales are more sensitive to intentional distortion than are other substantive personality scales (Hough et al., 1990; Stanush, 1997). Meta-analytic research has indicated that impression management scales are somewhat better than self-deceptive enhancement scales in detecting intentional distortion (Stanush, 1997).

Although such lie scales appear to be sensitive to intentional distortion, other ways of evaluating the usefulness of lie scales can lead to a different conclusion. For example, when lie scales are used in moderated regression analyses to determine whether criterion-related validity is affected by intentional distortion, the results indicate that intentional distortion does not affect validity, at least as measured with traditional lie scales (e.g., unlikely-virtues scales). Similarly, when traditional lie scale scores are used to adjust or correct other personality scale scores, validities are unaffected and corrected and raw scale scores are similar. Indeed, a large body of research has shown that traditional unlikely-virtues scales do not moderate, suppress, or mediate personality–criterion relationships, and adjusting faked personality scores for social desirability does

not undo the psychometric consequences of faking (Ellingson et al., 1999, 2001; Hough, 1998; Hough et al., 1990; Li & Bagger, 2006; McGrath, Mitchell, Kim, & Hough, 2010; Moorman & Podsakoff, 1992; Ones et al., 1996; Schmitt & Oswald, 2006). These data appear to suggest that traditional lie scales are ineffective in detecting intentional distortion. However, this assumes that (a) intentional distortion affects validity in deleterious ways and (b) the samples with which the analyses were conducted had a sufficient base rate of intentional distortion to test the hypotheses.

Studies that have found that traditional social desirability scales do not moderate, suppress, or mediate personality–criterion relationships (e.g., Hough et al., 1990; Ones et al., 1996) are based on research with incumbents and job applicants that is not characterized by significant distortion such as in directed-faking studies. The conclusion that traditional social desirability scales are irrevocably flawed may be premature.

Historically, an important research finding about traditional social desirability scales is that although they may measure response bias, they nonetheless correlate to a small degree with other personality traits and criteria of interest. Thus, adjusting personality scores based on traditional social desirability scales may remove valid trait variance from measures more than it adjusts for distortion, although research has indicated that criterion-related validity is unaffected, at least in samples in which significant faking is not present.

An important finding in this research stream is that social desirability is often nonlinearly associated with trait level (Kuncel & Tellegen, 2009). People do not always view the highest options as the most desirable options, either at a trait level or for the purposes of faking. Thus, simple linear corrections are not likely to correct socially desirable responding appropriately.

Current wisdom about traditionally developed unlikely-virtues scales is that they should not be used to make decisions about applicants (Dilchert & Ones, 2012). That is, respondents should not be singled out on the basis of their unlikely-virtues scale scores and informed that they produced invalid self-descriptions. Nor should scores on substantive personality scales be corrected or adjusted on the basis of scores on traditionally developed unlikely-virtues scales. More nuanced scaling of unlikely-virtues scale items may lead to a different conclusion. More research is needed.

Alternate approaches to assessing faking exist. In general, these approaches create scales based on more nuanced differences between item-response patterns among honest versus faked response sets. For example, a conceptually sound approach uses idiosyncratic item-response pattern scoring to detect intentional distortion. Kuncel and Borneman (2007) created a faking detection scale from faked versus honest differences in response-option endorsement patterns. The scale detected deliberate faking and was uncorrelated with valid trait variance. Equally important, these analyses were performed on samples characterized by significant (directed) faking.

What are the consequences of faking? Directed-faking studies have indicated that the validity, both construct and criterion related, of personality scales is diminished significantly when respondents are directed to distort their self-descriptions. In directed-faking studies, when participants are instructed to fake, multicollinearity increases and personality variables lose their construct validity (Ellingson et al., 2001; Schmit & Ryan, 1993; Stanush, 1997; Zickar & Robie, 1999). Distinct personality characteristics collapse into fewer factors or just one factor, losing their convergent and discriminant validity. Criterion validity is similarly significantly diminished. In directed-faking studies, meta-analytically obtained personality–performance correlations that were on average .35 in honest responding situations drop to .09 when respondents are directed to fake (Stanush, 1997). Clearly, when people are involved in experimental studies in which they are instructed to distort their self-descriptions, construct and criterion-related validity are markedly and significantly lower than when honest self-descriptions are provided. Important additional questions are "Do real-life job applicants fake?" "If yes, to what extent?" and "Do conclusions from directed-faking experiments generalize to real-life applicant settings?"

Do real-life job applicants fake? Researchers and practitioners agree that many job applicants do distort their answers to portray themselves in a more favorable light. The extent of applicant faking, however, is a point of disagreement among researchers; it appears to depend on several factors.

Donovan, Dwight, and Hurtz (2003) surveyed college students who had recently applied for a job; 32% claimed to have exaggerated their positive characteristics and downplayed their negative characteristics on a personality test (the distortion was apparently more exaggeration than outright falsification). Although this study provided information about applicants' self-reported distortion, it did directly answer the question about the amount of actual distortion in applicant settings. Moreover, research has indicated that when respondents are warned that a lie scale can detect distortion and that responses will be verified, intentions to fake are significantly reduced (Chen, Lee, & Yen, 2004).

Other research has documented the spread within an organization over time of extremely favorable responding to personality tests included in the organization's managerial promotion test battery, resulting in substantially higher test scores for test takers. Internal candidates who had initially failed the screen and were retested scored on average higher than the main group. More important, the organization effectively deterred distortion when it implemented a warning against distortion (Landers, Sackett, & Tuzinski, 2011).

Other forms of evidence are also relevant. Meta-analytic evidence comparing applicant personality scale mean scores to incumbent mean scores has indicated that applicants' mean scores tend to be between 0.10 and 0.50 standard deviation higher those of than incumbents (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). Of course, such between-groups designs are only informative when there are no differences in true personality means between applicant and incumbent groups and comparisons are appropriate. Often, studies that use between-group designs have not even matched applicant and incumbent groups on jobs (Birkeland et al., 2006).

Large-scale studies involving applicants to and incumbents in the same organization and doing the same type of work help shed light on the issue. In three very large real-life applicant and incumbent (honest) samples in a study involving mean-score comparisons between applicants for and incumbents in the same job in the same organization, mean-score differences between applicants and incumbents ranged from 0.04 to 0.56 standard deviation on seven different personality scales, with an average difference of about 0.25 standard deviation (Hough, 1998). In this study, applicant sample sizes were 25,423 for the police jobs, 14,442 for the telecommunications jobs, and 681 for state trooper jobs; incumbent samples were 508, 963, and 270, respectively. Both applicant and incumbent samples in this real-life study were sufficiently large to provide reliable estimates of the difference between honest respondents (incumbents) and motivated-to-distort respondents (job applicants in high-stakes testing situations).

A handful of studies have examined within-person change from applicant to incumbent contexts; the results have been mixed. Some studies found marked increases in scores for applicants, whereas others did not.

Studies examining validity, both construct and criterion related, have also been very revealing. First, applicants' personality scores do not typically show the same collapsed factor structure observed in directed-faking studies (e.g., Montag & Comrey, 1990; Schmit & Ryan, 1993). Instead, construct validity, a variable's convergent and discriminant validity, remains reasonably well intact. Second, criterion-related validities remain at useful levels.

A very large meta-analysis of criterion-related validity coefficients for integrity tests included personality-based integrity tests. Summarizing across 62 studies and 93,092 applicants, Ones et al. (1993) found a predictive criterion-related validity of .29. Hough (1998) examined the predictive criterion-related validities of personality variables obtained in applicant settings (respondents motivated to present themselves favorably) and the concurrent criterion-related validities obtained in incumbent settings (honest self-descriptions), concluding that setting does indeed moderate relationships between personality variables and criterion variables, but not in the way that many might expect.

Yes, in settings such as directed-faking studies, personality scores increase, and yes, personality–performance relationships in directed-faking studies are seriously compromised. However, in real-life personnel selection settings, faking does not appear to be that common, and personality–performance relationships are only minimally smaller than what are found in concurrent validation studies with incumbents. The rank order of actual job candidates in real-life hiring situations does not change to the extent of that of participants in directed-faking studies. Conclusions about the usefulness of personality variables based on directed-faking studies do not necessarily generalize to applicant settings.

Finding comparable criterion-related validity, comparable factor structures, and score stability among applicants offers some reassurance that applicant settings do not generally undermine personality's validity. Taken together, research has indicated that although some intentional distortion can and likely does occur in real-life applicant settings, it is not necessarily widespread, and personality–performance relationships generally remain intact.

Correlation coefficients are relatively robust to substantial changes in small portions of the sample, particularly when those changes are relevant to only a certain portion of the distribution (as might be expected among fakers; Rosse, Stecher, Miller, & Levin, 1998). Thus, even when faking is relatively infrequent, fakers may move to the top of the distribution and be more likely to be hired. Whether their job performance is lower (e.g., Mueller-Hanson, Heggestad, & Thornto, 2003) or higher (e.g., Hough, 1998) is unclear. It is clear, however, that faking is more problematic in low selection ratio settings (settings in which only a small portion of the applicant group is selected) than in high selection ratio settings.

Coaching is another factor to consider in real-life personnel selection. According to White, Young, Hunter, and Rumsey (2008), intentional distortion is an issue for the U.S. Army, particularly because recruiters often coach applicants on how to answer the personality items to ensure they pass the screen.

Employers are always concerned about the possibility of intentional distortion and its potential for affecting the predictive accuracy of their hiring decisions. As a result, the quest to develop more robust measures of personality and to deter intentional distortion continues.

Can faking be deterred? Several strategies have been proposed and examined, and for some the results are encouraging. Types of strategies include test administration strategies and test development strategies. Test administration strategies and their effectiveness are described in this section. (Test development strategies to overcome intentional distortion and provide more accurate self-descriptions are addressed in the Test Development Efforts to Overcome Intentional Distortion section.)

Warnings and consequences. Warning applicants both that faking can be detected and that there are consequences for doing so reduces distortion, although warning applicants only that faking can be detected appears not to reduce distortion (Dwight & Donovan, 2003). Another meta-analysis found that warnings were very effective deterrents against distortion in real-life applicant settings but not in directed-faking studies (Stanush, 1997). Introducing a warning after significant and blatant intentional distortion had spread throughout a company's selection process appeared to deter distortion (Landers et al., 2011). A variety of warnings–consequences combinations exist. One of the more effective warnings–consequences combinations is informing respondents that their answers will be verified (Vasilopoulos, Cucina, & McElreath, 2005). Another effective warnings–consequences combination is to inform respondents that they will be asked to expand on and explain their answers in an interview (Doll, 1971).

Evidence is accumulating that the validity of personality variables remains intact when warnings not to distort are included in instructions. Examples of studies that found no improvement in criterion-related validity include Converse et al. (2008); Fox and Dinur (1988); and Robson, Jones, and Abraham (2007). McFarland (2003), however, found less multicollinearity among personality variables when respondents were warned, indicating greater discriminant validity. More important, when respondents were warned, their scores on personality

variables incremented criterion-related validity over and above validity obtained using just cognitive ability (Converse et al., 2008).

A limited amount of research exists on the effects of warnings on respondents' perception of fairness and procedural justice. Three studies found that warnings do not negatively affect perceptions of fairness (i.e., Dullaghan & Borman, 2009; McFarland, 2003; Robson et al., 2007). One study found limited support for somewhat more negative reactions by some applicants to warnings (Converse et al., 2008).

The combination warning of lie scales and consequences appears to have considerable merit with little, if any, negative impact on applicant perceptions. Nonetheless, a word of caution is appropriate. Books in the popular press instruct and coach individuals on how to fake applicant personality inventories (e.g., *Employment Personality Tests Decoded;* Hartley & Sheldon, 2007), with sections devoted to avoiding lie detection scales. These books are effective not only in coaching individuals to increase their scores on substantive (content) scales but also in avoiding high scores on social desirability scales (Wolford & Christiansen, 2008). The advice these books provide may give respondents the confidence to ignore warnings if they believe they can avoid being detected as having portrayed themselves in an overly virtuous way.

Written elaboration. The effect of requiring respondents to elaborate on their answers results in significantly lower scores on those items as well as on other items in the inventory (Ramsay, Schmitt, Oswald, Kim, & Gillespie, 2006; Schmitt & Kunce, 2002). Although items were verifiable biodata items, such items are often included in personality measures. Research with measures of facets of Conscientiousness has produced similar results (Dubin, 2011).

Grouped versus random item sequence. Some researchers have recommended grouping items and labeling them, informing respondents what personality characteristic is being measured (cf. Morgeson et al., 2007). However, measures of Emotional Stability and Conscientiousness are more easily distorted when the items measuring the constructs are grouped together (McFarland, Ryan, & Ellis, 2002).

Test Development Efforts to Overcome Intentional Distortion

In spite of the evidence that validity generally remains intact in real-life personnel selection settings, applicants clearly can distort their answers, and that reality concerns users of self-report inventories. This concern is the underlying motivation for many test development efforts. In this section, we describe several lines of research, all devoted to increasing the accuracy of predictions using self-report inventories.

Item characteristics of criterion-valid items. Researchers have examined the characteristics that enhance criterion-related validity, and two features appear to have merit: verifiable items and items that ask respondents to make comparative judgments (Mabe & West, 1982; Schmidt & Hunter, 1998). Yet, these two features appear to be at odds with each other. Take, for example, the item characteristic verifiability. An item that asks "How often do you exercise?" and provides options "more than 5 hours per week" and "5 hours or less per week" is a verifiable item. If the options were "more often than your friends," "about as often as your friends," and "not quite as often as your friends," the item is asking for a comparative judgment. Which item is better? Perhaps both items are equally good, and it is in comparison with response options such "Yes" or "No" that both the verifiable and comparison judgment items result in less distortion.

Forced-choice self-report measurement strategies. Interest in and research on forced-choice formats is experiencing a resurgence. Instead of asking respondents to describe themselves using a single stem (Likert-type item) such as "I enjoy intellectual games" with response options such as *strongly agree* and *strongly disagree,* forced-choice items present more diverse options and ask the respondent to choose the option that is most (or least) descriptive of them. An example of a forced-choice item is "Which is more descriptive of you? 'I enjoy intellectual games' or 'I complete assignments on time?'"

The typical multidimensional, forced-choice scale scores have undesirable characteristics. One important negative outcome is ipsativity in the scores. Choosing one stem means not choosing a

different stem, resulting in a higher score on one characteristic and a lower score on the other. That is, scale scores provide information about a person's trait level relative to his or her other traits, an intraperson comparison. They do not provide information about a person's trait level relative to other people (no interperson comparison or normative information). This is a serious drawback if decisions are made about people on the basis of how they score relative to other people.

Such is the situation with personnel selection, placement, and promotion decisions. Consider, for example, a scenario in which an employer intends to hire one person, the most conscientious person. The choice is between Person A, who is higher on Extraversion than Conscientiousness, and Person B, who is higher on Conscientiousness than Extraversion but lower on both traits than Person A. A selection decision based on a typical forced-choice format will produce a higher score for Person B on Conscientiousness. In this scenario, the less Conscientious candidate would be hired because forced-choice scales measure within-individual differences rather than between-individual differences.

The ipsative nature of forced-choice measures forces a pattern of negative correlations between personality scales that would not otherwise be observed (Hicks, 1970). The ipsativity also distorts personality–criterion relationships. These effects are lessened when a large number of scales are used or when respondents choose among a larger number of stems.

Advocates of forced-choice measures have argued that pairing equally desirable stems will effectively reduce susceptibility to faking (e.g., Gordon, 1951). This strategy was used to develop the Edwards Personal Preference Schedule (Edwards, 1954). However, items presented in a forced-choice format are not immune to intentional distortion. A qualitative review of a large number of studies of directed-faking on forced-choice scales concluded that respondents can successfully distort their self-descriptions when instructed to do so (Waters, 1965). A quantitative review (meta-analytic summary) of directed-faking studies found a similar result: Although distortion was somewhat weaker with forced-choice formats than with normative (Likert-type) formats, responses to forced-choice items are also easily distorted (Stanush, 1997).

Some research has found that forced-choice measures retain predictive validity even under directed-faking conditions (Christiansen, Burns, & Montgomery, 2005; Hirsh & Peterson, 2008). However, faking instructions substantially inflate correlations between forced-choice personality measures and cognitive ability (Christiansen et al., 2005; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006), a finding not observed with normative (Likert-type) measures. Thus, some of forced-choice measures' retention of validity may stem from that part of the measure that captures cognitive ability. Another concern is the effect of coaching. In high-stakes testing, it is relatively easy to coach or instruct applicants in how to choose response options that measure the characteristics that are important to a decision maker.

More recently, researchers have applied item response theory methods to scoring forced-choice measures (Chernyshenko et al., 2009; Stark, Chernyshenko, & Drasgow, 2005). Normal item response theory conceptualizes the probability of endorsing an item as a logistic function of an underlying latent trait. However, choosing a forced-choice response from among two options is a function of a person's standing on two different traits. Stark et al.'s (2005) multidimensional pairwise preference model posits the item response curves for the probability of choosing a particular response as a function of multiple underlying traits in multidimensional space. Such multidimensional representations of forced-choice items permit the recovery of normative information about individuals' trait standing from an ipsative response format. Score inflation in faking contexts on these measures is generally less than that on typical Likert-type measures (White et al., 2008), and multidimensional pairwise preference scoring does not produce negative intercorrelations among scales that typify other forced-choice measures (Chernyshenko et al., 2009). Although multidimensional pairwise preference scored measures showed strong validities for predicted self-reported behaviors (Chernyshenko et al., 2009), when pitted against rationally developed biodata inventory scales, criterion-related validity was as

high or higher for the rational biodata scales (Heffner & Owens, 2011). It remains to be seen whether faking instructions produce strong correlations with general mental ability using these scoring methods. Similarly, the effects of coaching on scale scores are needed. Nonetheless, this research stream has produced a strong method for recovering normative personality information from a method thought to be too flawed (e.g., Hicks, 1970) for most uses.

Subtle (disguised) item content. Items for which the trait that is measured is transparent are referred to as *obvious items*. *Subtle items* are at the other end of the transparency continuum. Recent research has indicated that adding a frame of reference to items increases the accuracy of prediction or criterion-related validity (Bing, Whanger, Davison, & VanHook, 2004; Hunthausen, Truxillo, Bauer, & Hammer, 2003; Lievens, De Corte, & Schollaert, 2008; Schmit, Ryan, Stierwalt, & Powell, 1995). The added frame of reference increases the items' transparency and provides respondents with a context in which to self-report. The measurement is more situational and more precise. It has become standard wisdom in I/O psychology that providing respondents with a context in which to provide their self-descriptions enhances validity and reliability. Yet, other recent evidence has suggested that more subtle items retain their validity in high-stakes applicant settings presumably because they are more difficult to distort (White et al., 2008). The issue remains unresolved.

Conditional reasoning measures. Although intending to measure personality characteristics, these measures are presented as reasoning tests, attempting to disguise the purpose of measurement. James and colleagues' (James, 1998; James, McIntyre, Glisson, Bowler, & Mitchell, 2004) development of conditional reasoning measures of aggression and achievement represents one such novel approach. This approach stems from the premise that undesirable traits are associated with ego-protective cognitive biases that individuals with these traits use to justify their actions. For example, aggressive individuals are more likely to perceive hostile intent in others (referred to as *hostile attribution bias*), making their aggressive behavior seemingly justifiable. Items present scenarios that prime these justification mechanisms. Although early estimates have suggested that the Conditional Reasoning Test of Aggression produced uncorrected validities substantially larger (mean $r = .44$; James et al., 2004) than those of self-report measures, more recent meta-analytic evidence has suggested more conservative but still useful validities (e.g., mean $r = .26$ for predicting counterproductive work behavior; Berry, Sackett, & Tobares, 2010). Conditional reasoning measures tend to be relatively uncorrelated with self-report measures, perhaps because of their implicit versus explicit nature.

COMPARABILITY OF SELF-REPORT MEASUREMENTS ACROSS TESTING MODES AND LANGUAGES AND CULTURES

The span of personality measurement at work has expanded considerably in recent decades, because of both the rise of globalization and the ease of administration afforded by the Internet. Thus, a body of research has emerged examining whether the responses provided in new languages, cultures, and methods of administration are comparable to those in which these measures were originally developed.

Mode of Testing

Technology has transformed the world of work, and employment testing is no exception. Today, large and small companies use both onsite and online computerized testing to screen applicants. Fortunately, considerable research has examined the equivalence of personality scale measurements administered onsite or online, including proctored and unproctored testing and paper-and-pencil testing. Other research, although limited, has compared the validity obtained and intentional distortion using the different test administration methods. An important issue for online administration of personality scales is the effect of retesting.

Score and structural equivalence. Studies have reported somewhat different results, although the preponderance of evidence has supported similar score and structural equivalence. Results have indicated that scores are similar in proctored and

unproctored settings (W. Arthur, Glaze, Villado, & Taylor, 2009, 2010; Griffith, Chmielowski, & Yoshita, 2007). However, when comparing scores obtained in a supervised (proctored) online testing situation with those obtained in an unsupervised (unproctored) online testing situation, structural models of personality provided better fit in the supervised testing situation (Oswald, Carr, & Schmidt, 2001).

Validity. Very few studies have examined the criterion-related validity of personality scores obtained from the different testing modes (Tippins, 2009). Two studies concluded that criterion-related validities are similar in the different testing modalities (Beaty et al., 2011; Chuah, Drasgow, & Roberts, 2006). One of the two studies (i.e., Beaty et al., 2011, which was done after Tippins's 2009 review) used meta-analytic techniques to analyze a database of 125 validity coefficients archived by a testing company, including both concurrent and predictive validity study designs as well as proctored onsite and unproctored online testing. Given the large number of validity coefficients involved, the results are likely stable.

Intentional distortion. Intentional distortion issues might, at first blush, be of less importance given the well-established finding that self-report personality measures are easily distorted. Assistance from others is unnecessary to distort self-report items. When test takers can, without assistance from others, distort their responses, it should not matter whether the testing setting is proctored or unproctored. Nonetheless, important knowledge has been gained by examining intentional distortion in the various testing situations.

Research has suggested that mean-score differences between honest (low-stakes) conditions and faking (high-stakes) conditions in unproctored tests delivered remotely are similar to those in proctored test administration settings (W. Arthur et al., 2009, 2010). Similarly, the magnitude of score elevation and percentage of individuals identified as providing overly virtuous self-descriptions are similar in the two conditions (Griffith et al., 2007). Another study with random assignment of participants to one of three testing conditions (paper and pencil, proctored computer lab, and unproctored online) used item response theory, mean-score difference, regression, and factor analyses to examine score and structural comparability in the three conditions (Chuah et al., 2006). They concluded that the three conditions produced equivalent scores.

A meta-analysis of intentional distortion in paper-and-pencil measures versus computerized measures has suggested that some variables moderate the amount of distortion in the two testing conditions (Richman, Kiesler, Weisband, & Drasgow, 1999). The ability to backtrack and change answers increases the score equivalence of computerized and paper-and-pencil measures, and anonymity also moderates the equivalence between the two testing modes. When developing a computer-administered personality inventory, an important software feature is the capability to backtrack and change responses, especially if scores will be compared with scores obtained on paper-and-pencil measures.

An interesting possibility with computerized testing is the use of latency measures to detect intentional distortion, which is especially relevant given the evidence that participants take longer to respond when trying to distort their responses (e.g., Chen et al., 2004; Holden, 1995; Vasilopoulos, Reilly, & Leaman, 2000; Vasilopoulos et al., 2005). However, it is also important to note that response time is correlated with familiarity with the applied-for job (Vasilopoulos et al., 2000). Moreover, response latency is conceptually and empirically related to personality characteristics. For example, people who are low on Extraversion are likely to respond more slowly than people high on Extraversion. Research evidence has confirmed the conceptual relationship (Eysenck, 1967). People lower on Emotional Stability take more time than those higher on the trait (Furnham, Forde, & Cotter, 1998). Conceptually, people who are high on attention to detail (Conscientiousness) are likely to respond more slowly than people low on attention to detail. In addition, coaching and practice may change speed of response. Thus, although computerized testing provides an opportunity to examine additional ways of detecting intentional distortion, the process will not be a straightforward endeavor, and success will be difficult to achieve.

A less complicated approach is to deter intentional distortion with intermittent prompts that warn respondents about the consequences of intentional distortion (W. Arthur & Glaze, 2011). Such prompts could be in response to extreme or unusual responses, although the effect of such messages on honest responders needs to be examined. At least three studies examined online warnings with real job applicants. One study included online warnings delivered to applicants in real time on the basis of their responses to items (Evans & Waldo, 2009). They found that warnings affected scores on skepticism more than any other characteristic (they appeared less trusting after being warned), followed by social confidence, detail orientation, and good impression. Need for control, goal orientation, and need to nurture were virtually unchanged, although standard deviations were not provided and effect sizes were not reported. A second study (Melcher, 2009) also examined the effects of warnings (combination warning–consequence) as well as onsite proctored and offsite unproctored conditions. The sample included 1,745 applicants from 19 companies. The combination warning–consequence condition required each applicant to read and agree to an "integrity agreement" acknowledging that he or she understood that "falsifications or misrepresentations of any kind will be considered just cause for rejection of the assessment or dismissal from employment, or employment-related opportunities" (Melcher, 2009, p. 5). Applicants in this condition were also required to identify which of two boxes correctly summarized the warning. Melcher (2009) found that compared with the no-warning condition, the warning–consequence condition produced statistically significant lower scores on the lie scale. She also found, in contrast to the results obtained in other studies described earlier, that the offsite, unproctored condition produced statistically significant lower scores on the lie scale than the onsite, proctored condition. A third study examining job applicant online intentional distortion also found statistically significant lower lie scale scores in a combination warning–consequences condition than in a no-warning condition (Chen et al., 2004).

Retesting. An important issue for online (Internet) testing is the possibility of retesting, potentially a significant amount of retesting with the risk of overexposure to item content (Tippins, 2009; Tippins et al., 2006). With the ease and accessibility of online administration, frequent retesting is a serious issue for which no straightforward or easy solution exists, at least not currently.

Comparability Across Languages and Cultures

Multinational corporations hire people from all over the world. If they are to use personality inventories as part of the selection process, issues similar to those explored in the Mode of Testing section are relevant. The work required to answer similar issues for dozens of languages and cultures is daunting if not impossible.

Many personality inventories have been developed in one language and translated into other languages. It is difficult but not impossible to do so and obtain similar score meanings, structures, and criterion-related validities with different translations. A few of the issues include construct equivalence, which is hampered by the availability of other personality inventories in different languages, relevant norm groups, differences in interpretation of a behavior, and differences in response style (Nyfield & Baron, 2000). The International Test Commission has published a set of standards for translating tests developed in one language into another language (Hambleton, 1999).

At least two lines of research have indicated that personality measures operate differently in different countries and languages. Studies examining the structure of personality in different countries and cultures have not necessarily "discovered" the Big Five personality factors (e.g., Ashton et al., 2004). Studies using differential item functioning analyses have concluded that items operate differently in different languages and cultures (Ellis, 1989; Ryan, Horvath, Ployhart, Schmitt, & Slade, 2000).

One of the more ambitious projects in this area developed a set of personality scales intended specifically for cross-cultural use. Schmit, Kilm, and Robie (2000) involved psychologists from many different cultures, languages, and countries in all phases of test development and validation. Evidence has indicated that construct measurement is similar across

several languages and cultures. Test development strategies such as these clearly can produce comparable measurements useful in several cultures. Nonetheless, the laborious process of translating the items into the different languages such that scores retain their equivalence across the languages is required.

ALTERNATIVES TO TRADITIONAL SELF-REPORT TRAIT MEASURES

Although personality research and practice in I/O psychology has been dominated by the use of general Likert-type or true–false self-report measures, emerging research has begun exploring the promise held in alternative approaches to measuring personality. We discuss these in the sections that follow, describing alternatives that measure personality from a different rating source (e.g., peer reports or interviewers) or that capitalize on situational variability in the expression of personality traits.

Others' Ratings

One simple variation on traditional, Likert-type personality measures is to solicit personality ratings from nonself sources. Specifically, peers (e.g., coworkers) can be asked to describe a target individual's personality by completing an inventory in which first-person pronouns have been replaced with third-person pronouns. Although such approaches are common in basic personality research, which has found that well-acquainted observers are quite accurate in rating targets (Connelly & Ones, 2010; Funder, 1995), use of observer reports has been scant in I/O psychology. Nonetheless, the empirical support for measuring personality via observers' ratings is strong. Tupes (1957, 1959), for example, showed that such ratings predict performance of Air Force officers 6 months later. Mount, Barrick, and Strauss (1994) further showed that observer ratings predict performance more strongly than do self-reports, with these findings holding across two small-scale meta-analyses (Connelly & Ones, 2010; Oh, Wang, & Mount, 2011). Sample findings included that the operational validity of a single observer's reports of Conscientiousness is $\rho_{ov} = .29$ (vs. $\rho_{ov} = .21$ for self-reports; Barrick, Mount, & Judge, 2001); an optimally weighted composite of the Big Five yields $R_{ov} = .38$ (vs. $R_{ov} = .27$ for self-reports). The stronger predictive validity of observer reports holds even for observers who are acquainted with targets outside the workplace (Connelly & Hülsheger, 2012) as well as for predicting a wide range of behaviors critical to leading a happy and healthy life (Connelly & Ones, 2010; Fiedler, Oltmanns, & Turkheimer, 2004; Smith et al., 2008). These findings suggest that observers may have a more accurate view of targets than targets have of themselves. Moreover, observer ratings offer the potential to collect personality ratings from multiple observers. Combining multiple raters could allow predictive validities to climb as high as $\rho = .55$ for Conscientiousness. Validities of this magnitude considerably exceed any previous estimates for personality and suggest that using a single rater—and especially a self-rater—has substantially underestimated the importance of personality in the workplace.

Although these findings are encouraging, they also raise numerous pragmatic questions about how observers' reports might be obtained in personnel selection settings. Specifically, no studies have examined observer reports in a selection context. Thus, whether observers' reports are more or less susceptible to faking and to adverse impact (particularly from stereotypes and prejudices potentially held by raters) remains unclear. Further research addressing these pragmatic issues for selection is needed. Although use of observers' ratings currently appears nonexistent for personnel selection, two inventories—the Campbell Leadership Index (D. P. Campbell, 1991) and the Leadership Multi-Rater Assessment of Personality (Warren, 2008)—use observers' reports of personality for developmental feedback in a 360-degree-feedback type of approach. The stronger validity for observers' reports compared with self-reports suggests that observers may be a more useful and novel source of personality information for development.

Beyond measuring personality using knowledgeable observers (i.e., observers well acquainted with targets), interest is emerging in measuring personality traits with interviews. Researchers have long known that interview ratings are positively influenced

by applicants' Emotional Stability, Extraversion, Openness, Conscientiousness, and Agreeableness (Huffcutt, Conway, Roth, & Stone, 2001; Salgado & Moscoso, 2002). However, researchers have more recently explored organizing the content of structured interviews to measure personality traits specifically. That is, rather than conceptualizing personality as an underlying influence on interview performance, interviews can represent a viable method of measuring personality traits. Measuring personality in interviews has long been a staple of diagnosis in clinical psychology, even for measuring "normal" personality traits (e.g., the Structured Interview for the Five-Factor Model; Trull et al., 1998). Although strangers are generally poor judges of targets' personality (Connelly & Ones, 2010), interviewers tend to be more accurate than strangers (Barrick, Patton, & Haugland, 2000), suggesting that they may be good judges of personality and particularly adept at detecting personality-related information. More important, Van Iddekinge, Raymark, and Roth (2005) found that even under directed-faking conditions, interview-based measures of personality did not show mean inflation or the collapsed factor structure that is common among self-report measures. These findings are certainly encouraging, but further research is needed comparing the validity of interview-based measures of applicants' personality to that of self-report measures and general job suitability–focused interviews.

Personality Measurement Capitalizing on Situational Variability

In response to the classic person–situation debate, personality psychologists have amassed an overwhelming amount of evidence indicating that personality traits endure over time and are expressed consistently across situations (Kenrick & Funder, 1988). Although this cross-situational stability is generally recognized, increasing research attention has turned toward studying the variability in trait expression across contexts (for a wide variety of contemporary viewpoints of person–situation interactionism, see Donnellan, Lucas, & Fleeson, 2009). In the sections that follow we discuss ways that I/O psychology has begun to study situational variability, either by measuring it directly (via experience sampling methodology) or by asking respondents to consider only specific contexts (via frame of reference measures).

Experience sampling. Experience-sampling research has found that not only is there considerable intraindividual variability in behavioral expression but there are also stable interindividual differences in intraindividual variability (Borkenau & Ostendorf, 1998; Fleeson, 2001; Fournier, Moskowitz, & Zuroff, 2008). More and more, personality researchers are studying stable differential personality expression across situations as a function of social roles (Roberts, 2007; Wood & Roberts, 2006), the characteristics of interaction partners (Andersen & Thorpe, 2009; Fournier et al., 2008; Fournier, Moskowitz, & Zuroff, 2009), and the trait-activating characteristics of situations (Tett & Burnett, 2003). These findings raise two important questions for studying and measuring personality at work: (a) How is the work context a (measurably) unique context for personality expression and (b) what implications does this variability in trait expression have for studying personality's effects on work behaviors and attitudes?

Frame of reference. Frame-of-reference personality measures capitalize on the unique contextual expression of personality at work; they adjust typical self-report personality items by adding an *at work* suffix to the end of each item. Whereas a respondent may consider a variety of potential contexts when completing general personality measures, frame-of-reference items direct respondents to consider personality expression only in the context related to the criterion of interest (Schmit et al., 1995). Frame-of-reference measures produce stronger correlations with performance outcomes than do self-report measures (Bing et al., 2004; Schmit et al., 1995), and this effect is not driven only by increased reliability (Lievens et al., 2008). Parallel findings have been observed for predicting domain-specific satisfaction (Heller, Ferris, Brown, & Watson, 2009; Slatcher & Vazire, 2009). However, frame-of-reference measures (a) tend to correlate near the limits of their reliability with general measures; (b) do not correlate any more strongly with coworkers' ratings of traits than do general measures (Small & Diefendorff, 2006); and (c) have been researched using a concurrent

validity research design. As previously described, predictive validity appears to suffer when item content is obvious, and providing a frame of reference is intended to reduce ambiguity. Although providing the respondent with a context for the behavior might increase concurrent validity, the effect on predictive validity is unknown. Further research is needed on what mechanisms and frame-of-reference measures work through and on the boundary conditions of their effectiveness.

Intraindividual variability. Few studies have directly examined workplace outcomes associated with intraindividual variability in trait expression. Minbashian, Wood, and Beckmann (2010), using experience sampling methodology, found that the workplace expression of conscientiousness varied predictably when momentary task demands were high, although people high in trait conscientiousness were less dependent on task demands than people low in trait conscientiousness. Heller, Weinblatt, and Rachman-Engel (2010) examined whether individuals were less satisfied at work when they engaged in behaviors inconsistent with their general level of extraversion. Interestingly, extraverts were less satisfied when exhibiting introverted behaviors, but there was no parallel effect for introverts. Future work should also explore whether state expression of traits is more or less effective when it is consistent with general trait standing. In addition, what effect greater intraindividual variability in trait expression has on performance remains unclear. Arguments for the importance of authenticity suggest that intraindividual variability in trait expression would have negative performance consequences, but the recent emphasis on adaptability would favor greater intraindividual variability. Although it is unlikely that this experience sampling methodology is feasible for selection purposes, further research could point toward useful and interesting developmental applications aimed at understanding intraindividual variability and situational contingencies of trait expression.

CONCLUSIONS

A vast amount of basic and applied research has been devoted to understanding the role of personality in the workplace. Knowledge has increased dramatically even in the past 10 years. Among the findings are the following:

- The five-factor model has been a very important organizing framework for understanding the influence of personality on important work criteria.
- The six-factor HEXACO model is an even better framework.
- Other strategies for developing a structure of personality variables, such as the nomological-web clustering approach, may provide greater insights for researchers seeking to understand the determinants of work performance and work adjustment criteria.
- Personality constructs predict many critically important workplace criteria, including overall job performance, contextual performance such as organizational citizenship and dedication, counterproductive work behavior such as disciplinary problems and absenteeism, managerial effectiveness, workplace safety, tenure and turnover, and team performance.
- Relationships between personality constructs and criterion constructs are logical; theoretical relevance of the predictor for the criteria is a major factor in building good models of the determinants of workplace performance and adjustment. Relationships are often indirect, moderated, and nonlinear.
- Measures of personality variables are useful for incrementing criterion-related validity over and above cognitive ability measures.
- Whites, African Americans, Hispanics, and men and women score similarly on most measures of personality variables; personality measures included in personnel selection procedures typically have little or no adverse impact on protected classes, although attending to mean-score differences at the facet level is important to ensure the least amount of difference.
- Self-report Likert-type measures are susceptible to intentional distortion.
- Nonetheless, in real-life applicant settings, criterion-related validities of self-report Likert-type measures remain essentially intact.

- Self-report forced-choice (non-Likert) measures of personality, often thought to be immune to response distortion, are susceptible to distortion but to a lesser degree than self-report Likert-type measures.
- Mode of measurement—paper and pencil, online proctored, and online unproctored—has little or no effect on scores or validity of personality measures.
- The most effective strategy for reducing overly virtuous responding in high-stakes testing for all of these testing strategies and testing modes is a combination warning–consequence message. Consequences, such as informing respondents that (a) they will need to explain their answers in an interview or in writing; (b) their answers will be verified; or (c) they will be eliminated from the hiring process, all appear to reduce intentional distortion, although some likely have other unintended consequences as well.
- Less distortion occurs on verifiable items such as biodata items.
- Others' reports of the target person's personality provide scores on personality measures that correlate even higher with important workplace criteria, although they too have issues of intentional distortion, especially descriptions obtained by family, parents, and close friends, who in research settings provide some of the most accurate descriptions of the target person.
- Situations do matter, and incorporating them into measurements may increase personality–criterion correlations, but providing situational context in measurements may decrease the personality–criterion correlations obtained in high-stakes testing situations, although this effect may be countered by including effective warning–consequence messages.

Many questions remain unanswered, many of which can only be answered with more sophisticated research designs and more nuanced hypotheses. There is much yet to learn.

References

Andersen, S. M., & Thorpe, J. S. (2009). An IF-THEN theory of personality: Significant others and the relational self. *Journal of Research in Personality, 43*, 163–170. doi:10.1016/j.jrp.2008.12.040

Arthur, J., Jr., Barrett, G. V., & Alexander, R. A. (1991). Prediction of vehicular accident involvement: A meta-analysis. *Human Performance, 4*, 89–105. doi:10.1207/s15327043hup0402_1

Arthur, W., Jr., & Glaze, R. M. (2011). Cheating and response distortion on remotely delivered assessments. In N. T. Tippins & S. Adler (Eds.), *Technology-enhanced assessment of talent* (pp. 99–152). San Francisco, CA: Jossey-Bass. doi:10.1002/9781118256022.ch4

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored Internet-based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology: Perspective on Science and Practice, 2*, 39–45. doi:10.1111/j.1754-9434.2008.01105.x

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored and Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*, 1–16. doi:10.1111/j.1468-2389.2010.00476.x

Arvey, R. D., Bouchard, T. L., Jr., Carroll, J. B., Cattell, R. B., Cohen, D. B., Dawis, R. V., . . . Willerman, L. (1994, December 13). Mainstream science on intelligence [Editorial]. *Wall Street Journal,* p. 356.

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19*, 289–303. doi:10.1002/(SICI)1099-1379(199805)19:3<289::AID-JOB841>3.0.CO;2-C

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., . . . De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*, 356–366. doi:10.1037/0022-3514.86.2.356

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta analysis. *Personnel Psychology, 44*, 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*, 9–30. doi:10.1111/1468-2389.00160

Barrick, M. R., Patton, G. K., & Haugland, S. N. (2000). Accuracy of interviewer judgments of job applicant personality traits. *Personnel Psychology, 53*, 925–951. doi:10.1111/j.1744-6570.2000.tb02424.x

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation.

Journal of Applied Psychology, 90, 1185–1203. doi:10.1037/0021-9010.90.6.1185

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92*, 410–424. doi:10.1037/0021-9010.92.2.410

Berry, C. M., Sackett, P. R., & Tobares, V. (2010). A meta-analysis of conditional reasoning tests of aggression. *Personnel Psychology, 63*, 361–384. doi:10.1111/j.1744-6570.2010.01173.x

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89*, 150–157. doi:10.1037/0021-9010.89.1.150

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317–335. doi:10.1111/j.1468-2389.2006.00354.x

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of Applied Psychology, 89*, 901–910. doi:10.1037/0021-9010.89.5.901

Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality, 32,* 202–221. doi:10.1006/jrpe.1997.2206

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment, 9*, 52–69. doi:10.1111/1468-2389.00163

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76*, 863–872. doi:10.1037/0021-9010.76.6.863

Campbell, D. P. (1991). *Manual for the Campbell Leadership Index*. Minneapolis, MN: National Computer Systems.

Campbell, J. P., & Pritchard, R. D. (1976). Motivation theory in industrial and organizational psychology. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 63–130). Chicago, IL: Rand McNally College.

Chan, D. (2009). So why ask me? Are self-report data really that bad? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences* (pp. 309–335). New York, NY: Routledge.

Chen, C. I., Lee, M. N., & Yen, C. L. (2004). Faking intention on the Internet: Effects of test types and situational factors. *Chinese Journal of Psychology, 46*, 349–359.

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*, 105–127. doi:10.1080/08959280902743303

Christian, M. S., Bradley, J. C., Wallace, J. C., & Burke, M. J. (2009). Workplace safety: A meta-analysis of the roles of person and situation factors. *Journal of Applied Psychology, 94*, 1103–1127. doi:10.1037/a0016172

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267–307. doi:10.1207/s15327043hup1803_4

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality, 40*, 359–376. doi:10.1016/j.jrp.2005.01.006

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*, 678–707. doi:10.1037/0021-9010.85.5.678

Connelly, B. S., & Hülsheger, U. R. (2012). A narrower scope or a clearer lens? Examining the validity of personality ratings from observers outside the workplace. *Journal of Personality*, *80*, 603–631. doi:10.1111/j.1467-6494.2011.00744.x

Connelly, B. S., & Ones, D. S. (2010). Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092–1122. doi:10.1037/a0021212

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 155–169. doi:10.1111/j.1468-2389.2008.00420.x

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment: An individual-differences model and its applications*. Minneapolis: University of Minnesota Press.

De Fruyt, F., & Mervielde, I. (1997). The five-factor model of personality and Holland's RIASEC interest types. *Personality and Individual Differences, 23*, 87–103. doi:10.1016/S0191-8869(97)00004-4

De Fruyt, F., & Mervielde, I. (1999). RIASEC types and Big Five traits as predictors of employment status and nature of employment. *Personnel Psychology, 52*, 701–727. doi:10.1111/j.1744-6570.1999.tb00177.x

Derue, D. S., Nahrgang, J. D., Wellman, N., & Humphrey, S. E. (2011). Trait and behavioral theories of leadership: An integration and meta-analytic test of their relative validity. *Personnel Psychology, 64*, 7–52. doi:10.1111/j.1744-6570.2010.01201.x

Dilchert, S., & Ones, D. S. (2008). Personality and extrinsic career success: Predicting managerial salary at different organizational levels. *Zeitschrift für Personalpsychologie, 7*, 1–23.

Dilchert, S., & Ones, D. S. (2012). Application of preventative strategies. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 177–200). New York, NY: Oxford University Press.

Doll, R. E. (1971). Item susceptibility to attempted faking as related to item characteristics and adopted fake set. *Journal of Psychology: Interdisciplinary and Applied, 77*, 9–16. doi:10.1080/00223980.1971.9916848

Donnellan, M. B., Lucas, R. E., & Fleeson, W. (2009). Personality and assessment at age 40: Reflections on the past person–situation debate and emerging directions of future person–situation integrations [Special issue]. *Journal of Research in Personality, 43*(2).

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*, 81–106. doi:10.1207/S15327043HUP1601_4

Dubin, D. (2011, April). *Can you elaborate? A novel approach for mitigating the effects of personality faking*. Poster presented at the 26th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57. doi:10.1037/0021-9010.91.1.40

Dullaghan, T. R., & Borman, W. C. (2009, April). *Effect of warnings against faking on personality tests for selection*. Poster presented at the 24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23. doi:10.1207/S15327043HUP1601_1

Edwards, A. L. (1954). *Edwards Personal Preference Schedule manual*. New York, NY: Psychological Corporation.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155–166. doi:10.1037/0021-9010.84.2.155

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86*, 122–133. doi:10.1037/0021-9010.86.1.122

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74*, 912–921. doi:10.1037/0021-9010.74.6.912

Evans, A. L., & Waldo, D. (2009, April). *You've been warned.* Paper presented at the 24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Eysenck, H. (1967). *The biological bases of personality*. Springfield, IL: Charles C Thomas.

Feist, G. J. (1998). A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review, 2*, 290–309. doi:10.1207/s15327957pspr0204_5

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80*, 1011–1027. doi:10.1037/0022-3514.80.6.1011

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. *Personnel Psychology, 61*, 579–616. doi:10.1111/j.1744-6570.2008.00123.x

Fournier, M. A., Moskowitz, D. S., & Zuroff, D. C. (2008). Integrating dispositions, signatures, and the interpersonal domain. *Journal of Personality and Social Psychology, 94*, 531–545. doi:10.1037/0022-3514.94.3.531

Fournier, M. A., Moskowitz, D. S., & Zuroff, D. C. (2009). The interpersonal signature. *Journal of Research in Personality, 43*, 155–162. doi:10.1016/j.jrp.2009.01.023

Fox, S., & Dinur, Y. (1988). Validity of self-assessments: A field evaluation. *Personnel Psychology, 41*, 581–592. doi:10.1111/j.1744-6570.1988.tb00645.x

Frei, R. L., & McDaniel, M. A. (1998). Validity of customer service measures in personnel selection: A review of criterion and construct evidence. *Human Performance, 11*, 1–27. doi:10.1207/s15327043hup1101_1

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652–670. doi:10.1037/0033-295X.102.4.652

Furnham, A., Forde, L., & Cotter, T. (1998). Personality and test taking style. *Personality and Individual Differences, 24*, 19–23. doi:10.1016/S0191-8869(97)00141-4

Galton, F. (1884). Measurement of character. *Fortnightly Review, 36*, 179–185.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26–34. doi:10.1037/0003-066X.48.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldberg, L. R., Sweeney, D., Merenda, P. F., & Hughes, J. E., Jr. (1998). Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes. *Personality and Individual Differences, 24*, 393–403. doi:10.1016/S0191-8869(97)00110-4

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology, 35*, 407–412. doi:10.1037/h0058853

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*, 341–355. doi:10.1108/00483480710731310

Hansen, C. P. (1988). Personality characteristics of the accident involved employee. *Journal of Business and Psychology, 2,* 346–365. doi:10.1007/BF01013766

Hansen, C. P. (1989). A causal model of the relationship among accidents, personality, and cognitive factors. *Journal of Applied Psychology, 74*, 81–90. doi:10.1037/0021-9010.74.1.81

Hartley, A., & Sheldon, G. (2007). *Employment personality tests decoded.* Franklin Lakes, NJ: Career Press.

Heffner, T., & Owens, K. (2011, April). Predicting job performance from non-cognitive measures. In T. S. Heffner & L. White (Chairs), *Advancing personality assessment for selection.* Symposium conducted at the 26th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Heller, D., Ferris, D. L., Brown, D., & Watson, D. (2009). The influence of work personality on job satisfaction: Incremental validity and mediation effects. *Journal of Personality, 77*, 1051–1084. doi:10.1111/j.1467-6494.2009.00574.x

Heller, D., Weinblatt, N., & Rachman-Engel, H. (2010, August). The role of consistency in extraversion in employee well-being: An experience sampling study. In D. B. Zoogah (Chair), *Individual differences.* Symposium conducted at the annual meeting of the Academy of Management, Montreal, Quebec, Canada.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184. doi:10.1037/h0029780

Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Research in Personality, 42*, 1323–1333. doi:10.1016/j.jrp.2008.04.006

Hogan, J., Hogan, R. T., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology, 69*, 167–173. doi:10.1037/0021-9010.69.1.167

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100–112. doi:10.1037/0021-9010.88.1.100

Holden, R. R. (1995). Response latency detection of fakers on personnel tests. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 27*, 343–355. doi:10.1037/0008-400X.27.3.343

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Hooper, A. C. (2007). *Self-presentation on personality measures in lab and field settings: A meta-analysis.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Hough, L. M. (1989). Development of personality measures to supplement selection decisions. In B. J. Fallon, H. P. Pfister, & J. Brebner (Eds.), *Advances in industrial organizational psychology* (pp. 365–375). Amsterdam, the Netherlands: Elsevier Science.

Hough, L. M. (1992). The "Big Five" personality variables—Construct confusion: Description versus prediction. *Human Performance, 5*, 139–155.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 209–244.

Hough, L. M., & Dilchert, S. (2007). *Inventors, innovators, and their leaders: Selecting for conscientiousness will keep you "inside the box"* (SIOP Leading Edge Consortium). Kansas City, MO: Society for Industrial and Organizational Psychology.

Hough, L. M., & Dilchert, S. (2010). Personality: Its measurement and validity for employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 299–319). New York, NY: Taylor & Francis.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities [Monograph]. *Journal of Applied Psychology, 75*, 581–595. doi:10.1037/0021-9010.75.5.581

Hough, L. M., & Furnham, A. (2003). Importance and use of personality variables in work settings.

In I. B. Weiner (Ed.-in-Chief) & W. Borman, D. Illgen, & R. Klimoski (Vol. Eds.), *Comprehensive handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 131–169). New York, NY: Wiley.

Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *International handbook of work and organizational psychology* (pp. 233–277). London, England: Sage.

Hough, L. M., Ones, D. S., & Viswesvaran, C. (1998, April). Personality correlates of managerial performance constructs. In R. C. Page (Chair), *Personality determinants of managerial potential performance, progression and ascendancy*. Symposium conducted at the 13th annual convention of the Society of Industrial and Organizational Psychology, Dallas, Texas.

Hough, L. M., & Schneider, R. J. (1996). Personality traits, taxonomies, and applications in organizations. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 31–88). San Francisco, CA: Jossey-Bass.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*, 897–913. doi:10.1037/0021-9010.86.5.897

Hunthausen, J. M., Truxillo, D. M., Bauer, T. N., & Hammer, L. B. (2003). A field study of frame-of-reference effects on personality test validity. *Journal of Applied Psychology, 88*, 545–551. doi:10.1037/0021-9010.88.3.545

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–879. doi:10.1037/0021-9010.85.6.869

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1*, 131–163. doi:10.1177/109442819812001

James, L. R., McIntyre, M. D., Glisson, C. A., Bowler, J. L., & Mitchell, T. R. (2004). The conditional reasoning measurement system for aggression: An overview. *Human Performance, 17*, 271–295. doi:10.1207/s15327043hup1703_2

Johnson, J. W., & Hezlett, S. A. (2008). Modeling the influence of personality on individuals at work: A review and research agenda. In S. Cartwright & C. L. Cooper (Eds.), *Oxford handbook of personnel psychology* (pp. 59–93). Oxford, England: Oxford University Press. doi:10.1093/oxfordhb/9780199234738.003.0004

Jordan, M., Herriot, P., & Chalmers, C. (1991). Testing Schneider's ASA theory. *Applied Psychology, 40*, 47–53. doi:10.1111/j.1464-0597.1991.tb01357.x

Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—self esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology, 86*, 80–92. doi:10.1037/0021-9010.86.1.80

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*, 765–780. doi:10.1037/0021-9010.87.4.765

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*, 530–541. doi:10.1037/0021-9010.87.3.530

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology, 52*, 621–652. doi:10.1111/j.1744-6570.1999.tb00174.x

Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology, 87*, 797–807. doi:10.1037/0021-9010.87.4.797

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person–situation debate. *American Psychologist, 43*, 23–34. doi:10.1037/0003-066X.43.1.23

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment, 15*, 220–231. doi:10.1111/j.1468-2389.2007.00383.x

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*, 201–228. doi:10.1111/j.1744-6570.2009.01136.x

Kwong, J. Y. Y., & Cheung, F. M. (2003). Prediction of performance facets using specific personality traits in the Chinese context. *Journal of Vocational Behavior, 63*, 99–110. doi:10.1016/S0001-8791(02)00021-0

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*, 202–210. doi:10.1037/a0020375

Lee, K., Ashton, M. C., & deVries, R. E. (2005). Predicting workplace delinquency and integrity with the HEXACO and five-factor models of personality structure. *Human Performance, 18*, 179–197. doi:10.1207/s15327043hup1802_4

LePine, J. A., Colquitt, J. A., & Erez, A. (2000). Adaptability to changing task contexts: Effects of general cognitive ability, conscientiousness and

openness to experience. *Personnel Psychology, 53*, 563–593. doi:10.1111/j.1744-6570.2000.tb00214.x

LePine, J. A., Erez, A., & Johnson, D. E. (2002). The nature and dimensionality of organizational citizenship behavior: A critical review and meta-analysis. *Journal of Applied Psychology, 87*, 52–65. doi:10.1037/0021-9010.87.1.52

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*, 131–141. doi:10.1111/j.1468-2389.2006.00339.x

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*, 268–279. doi:10.1037/0021-9010.93.2.268

Lievens, F., Ones, D. S., & Dilchert, S. (2009). Personality scale validities increase throughout medical school. *Journal of Applied Psychology, 94*, 1514–1535. doi:10.1037/a0016137

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280–296. doi:10.1037/0021-9010.67.3.280

McFarland, L. A. (2003). Warning against faking on a personality test: Effects on applicant reactions and personality test scores. *International Journal of Selection and Assessment, 11*, 265–276. doi:10.1111/j.0965-075X.2003.00250.x

McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behavior and test measurement properties. *Journal of Personality Assessment, 78*, 348–369. doi:10.1207/S15327752JPA7802_09

McGrath, R. E., Mitchell, M., Kim, B., & Hough, L. M. (2010). The validity of response bias indicators. *Psychological Bulletin, 136*, 450–470. doi:10.1037/a0019216

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335–354. doi:10.1111/j.1744-6570.1990.tb01562.x

Melcher, K. M. (2009, April). Effectiveness of techniques for decreasing faking on unproctored Internet tests. In K. M. Melcher (Chair), *Remote assessment and applicant response distortion: Applied research and practice*. Symposium conducted at the 24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Minbashian, A., Wood, R. E., & Beckmann, N. (2010). Task-contingent conscientiousness as a unit of personality at work. *Journal of Applied Psychology, 95*, 793–806. doi:10.1037/a0020016

Mol, S. T., Born, M. P. H., Willemsen, M. E., & Van Der Molen, H. T. (2005). Predicting expatriate job performance for selection purposes: A quantitative review. *Journal of Cross-Cultural Psychology, 36*, 590–620. doi:10.1177/0022022105278544

Montag, I., & Comrey, A. L. (1990). Stability of major personality factors under changing motivational conditions. *Journal of Social Behavior and Personality, 5*, 265–274.

Moon, H., Hollenbeck, J. R., Marinova, S., & Humphrey, S. E. (2008). Beneath the surface: Uncovering the relationship between extraversion and organizational citizenship behavior through a facet approach. *International Journal of Selection and Assessment, 16*, 143–154. doi:10.1111/j.1468-2389.2008.00419.x

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behavior research. *Journal of Occupational and Organizational Psychology, 65*, 131–149. doi:10.1111/j.2044-8325.1992.tb00490.x

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729. doi:10.1111/j.1744-6570.2007.00089.x

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71–83. doi:10.1207/s15327043hup1002_1

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11*, 145–165.

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology, 79*, 272–280. doi:10.1037/0021-9010.79.2.272

Neuman, G. A., & Wright, J. (1999). Team effectiveness: Beyond skills and cognitive ability. *Journal of Applied Psychology, 84*, 376–389. doi:10.1037/0021-9010.84.3.376

Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology, 58*, 367–408. doi:10.1111/j.1744-6570.2005.00515.x

O'Brien, K. M., & Fassinger, R. E. (1993). A causal model of the career orientation and career choice of adolescent women. *Journal of Counseling Psychology, 40*, 456–469. doi:10.1037/0022-0167.40.4.456

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of

personality traits: A meta-analysis. *Journal of Applied Psychology, 96*, 762–773.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027. doi:10.1111/j.1744-6570.2007.00099.x

Ones, D. S., & Viswesvaran, C. (1998a). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245–269.

Ones, D. S., & Viswesvaran, C. (1998b). Integrity testing in organizations. In R. W. Griffin, A. O'Leary-Kelly, & J. M. Collins (Eds.), *Dysfunctional behavior in organizations: Vol. 2. Nonviolent behaviors in organizations* (pp. 243–276). Greenwich, CT: JAI Press.

Ones, D. S., & Viswesvaran, C. (2001). Personality at work: Criterion-focused occupational personality scales (COPS) used in personnel selection. In B. W. Roberts & R. T. Hogan (Eds.), *Applied personality psychology: The intersection of personality and I/O psychology* (pp. 63–92). Washington, DC: American Psychological Association.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660–679. doi:10.1037/0021-9010.81.6.660

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679–703. doi:10.1037/0021-9010.78.4.679

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2003). Personality and absenteeism: A meta-analysis of integrity tests. *European Journal of Personality, 17*(Suppl. 1), S19–S38. doi:10.1002/per.487

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology, 48*, 775–802. doi:10.1111/j.1744-6570.1995.tb01781.x

Orvis, K. A., Brusso, R. C., Wasserman, M. E., & Fisher, S. L. (2010). E-nabled for E-learning? The moderating role of personality in determining the optimal degree of learner control in an E-learning environment. *Human Performance, 24*, 60–78. doi:10.1080/08959285.2010.530633

Oswald, F. L., Carr, J. Z., & Schmidt, A. M. (Chairs). (2001, April). *The medium and the message: Dual effects of supervision and Web-based testing on measurement equivalence for ability and personality measures.* Symposium conducted at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Oswald, F. L., & Hough, L. M. (2011). Personality and its assessment in organizations: Theoretical and empirical developments. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Vol. 2. Selecting and developing members for the organization* (pp. 153–184). Washington, DC: American Psychological Association.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609. doi:10.1037/0022-3514.46.3.598

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology, 84*, 411–422. doi:10.1037/0022-3514.84.2.411

Paunonen, S. V., & Nicol, A. A. A. M. (2001). The personality hierarchy and the prediction of work behaviors. In B. W. Roberts & R. T. Hogan (Eds.), *Personality psychology in the workplace* (pp. 161–191). Washington, DC: American Psychological Association. doi:10.1037/10434-007

Peeters, M. A. G., Van Tuijl, H. F. J. M., Rutte, C. G., & Reymen, I. M. M. J. (2006). Personality and team performance: A meta-analysis. *European Journal of Personality, 20*, 377–396. doi:10.1002/per.588

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*, 322–338. doi:10.1037/a0014996

Rainey, L. M., & Borders, L. D. (1997). Influential factors in career orientation and career aspiration of early adolescent girls. *Journal of Counseling Psychology, 44*, 160–172. doi:10.1037/0022-0167.44.2.160

Rauch, A., & Frese, M. (2007). Let's put the person back into entrepreneurship research: A meta-analysis on the relationship between business owners' personality traits, business creation, and success. *European Journal of Work and Organizational Psychology, 16*, 353–385. doi:10.1080/13594320701595438

Roberts, B. W. (2007). Contextualizing personality psychology. *Journal of Personality, 75*, 1071–1082. doi:10.1111/j.1467-6494.2007.00467.x

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58*, 103–139. doi:10.1111/j.1744-6570.2005.00301.x

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*, 313–345. doi:10.1111/j.1745-6916.2007.00047.x

Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology, 66*, 225–244. doi:10.1111/j.2044-8325.1993.tb00534.x

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance, 21*, 89–106. doi:10.1080/08959280701522155

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644. doi:10.1037/0021-9010.83.4.634

Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., & Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology, 53*, 531–562. doi:10.1111/j.1744-6570.2000.tb00213.x

Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology, 82,* 30–43. doi:10.1037/0021-9010.82.1.30

Salgado, J. F., & Moscoso, S. (2002). Comprehensive metaanalysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11*, 299–324. doi:10.1080/13594320244000184

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implication of 85 years of research findings. *Psychological Bulletin, 124,* 262–274. doi:10.1037/0033-2909.124.2.262

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*, 432–439. doi:10.1037/0021-9010.71.3.432

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (1997). Validity of integrity tests for predicting drug and alcohol abuse: A meta-analysis. In W. J. Bukoski (Ed.), *Meta-analysis of drug abuse prevention programs* (pp. 69–95). Rockville, MD: National Institute on Drug Abuse.

Schmit, M. J., Kilm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology, 53*, 153–193. doi:10.1111/j.1744-6570.2000.tb00198.x

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*, 966–974. doi:10.1037/0021-9010.78.6.966

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80*, 607–620. doi:10.1037/0021-9010.80.5.607

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology, 91*, 613–621. doi:10.1037/0021-9010.91.3.613

Schneider, B., Smith, D. G., Taylor, S., & Fleenor, J. (1998). Personality and organizations: A test of the homogeneity of personality hypothesis. *Journal of Applied Psychology, 83,* 462–470. doi:10.1037/0021-9010.83.3.462

Schneider, P. L., Ryan, J. M., Tracey, T. J. G., & Rounds, J. (1996). Examining the relation between Holland's RIASEC model and the interpersonal circle. *Measurement and Evaluation in Counseling and Development, 29*, 123–133.

Schneider, R. J., Ackerman, P. L., & Kanfer, R. (1996). To "act wisely in human relations": Exploring the dimensions of social competence. *Personality and Individual Differences, 21*, 469–481. doi:10.1016/0191-8869(96)00084-0

Schneider, R. J., Hough, L. M., & Dunnette, M. D. (1996). Broadsided by broad traits: How to sink science in five dimensions or less. *Journal of Organizational Behavior, 17*, 639–655. doi:10.1002/(SICI)1099-1379(199611)17:6<639::AID-JOB3828>3.0.CO;2-9

Slatcher, R. B., & Vazire, S. (2009). Effects of global and contextualized personality on relationship satisfaction. *Journal of Research in Personality, 43*, 624–633. doi:10.1016/j.jrp.2009.02.012

Small, E. E., & Diefendorff, J. M. (2006). The impact of contextual self-ratings and observer ratings of personality on the personality-performance relationship. *Journal of Applied Social Psychology, 36*, 297–320. doi:10.1111/j.0021-9029.2006.00009.x

Smith, T. W., Uchino, B. N., Berg, C. A., Florsheim, P., Pearce, G., Hawkins, M., . . . Yoon, H.-C. (2008). Associations of self-reports versus spouse ratings of negative affectivity, dominance, and affiliation with coronary artery disease: Where should we look and who should we ask when studying personality and health? *Health Psychology, 27*, 676–684. doi:10.1037/0278-6133.27.6.676

Spector, P. E., Jex, S. M., & Chen, P. Y. (1995). Relations of incumbent affect-related personality traits with incumbent and objective measures of characteristics of jobs. *Journal of Organizational Behavior, 16*, 59–65. doi:10.1002/job.4030160108

Stanush, P. L. (1997). *Factors that influence the susceptibility of self-report inventories to distortion: A meta-analytic investigation*. Unpublished doctoral dissertation, Texas A&M University, College Station.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring

pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*, 184–203. doi:10.1177/0146621604273988

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin, 133*, 65–94. doi:10.1037/0033-2909.133.1.65

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500–517. doi:10.1037/0021-9010.88.3.500

Tett, R. P., Jackson, D. N., Rothstein, M., & Reddon, J. R. (1994). Meta-analysis of personality-job performance relations: A reply to Ones, Mount, Barrick, and Hunter (1994). *Personnel Psychology, 47*, 157–172. doi:10.1111/j.1744-6570.1994.tb02415.x

Thoresen, C. J., Kaplan, S. A., Barsky, A. P., de Chermont, K., & Warren, C. R. (2003). The affective underpinnings of job perceptions and attitudes: A meta-analytic review and integration. *Psychological Bulletin*, *129*, 914–945. doi:10.1037/0033-2909.129.6.914

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspective on Science and Practice, 2*, 2–10.

Tippins, N. T., Beatty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology*, *59*, 189–225. doi:10.1111/j.1744-6570.2006.00909.x

Trull, T. J., Widiger, T. A., Useda, J. D., Holcomb, J., Doom, B. T., Axelrod, S. R., . . . Gershuny, B. S. (1998). A structured interview for the assessment of the five-factor model of personality. *Psychological Assessment, 10*, 229–240. doi:10.1037/1040-3590.10.3.229

Tupes, E. C. (1957). *Relationships between behavior trait ratings by peers and later office performance of USAF Officer Candidate School graduates* (No. AFPTRC-TN-57–125). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Tupes, E. C. (1959). *Personality traits related to effectiveness of junior and senior Air Force officers* (WADC-TN59–198, AD-231 256). Lackland Air Force Base, TX: Wright Air Development Center.

Tupes, E. C., & Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality, 60*, 225–251. (Original work published 1961) doi:10.1111/j.1467-6494.1992.tb00973.x

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology, 90*, 536–552. doi:10.1037/0021-9010.90.3.536

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*, 175–199. doi:10.1207/s15327043hup1903_1

Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 90*, 306–322. doi:10.1037/0021-9010.90.2.306

Vasilopoulos, N. L., Reilly, R. R., & Leaman, J. A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85*, 50–64. doi:10.1037/0021-9010.85.1.50

Vinchur, A. J., Schippmann, J. S., Switzer, F. S., & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83*, 586–597. doi:10.1037/0021-9010.83.4.586

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210. doi:10.1177/00131649921969802

Wanberg, C. R., Hough, L. M., & Song, Z. (2002). Predictive validity of a multidisciplinary model of reemployment success. *Journal of Applied Psychology, 87*, 1100–1120. doi:10.1037/0021-9010.87.6.1100

Wanberg, C. R., Watt, J. D., & Rumsey, D. J. (1996). Individuals without jobs: An empirical study of job-seeking behavior and reemployment. *Journal of Applied Psychology, 81*, 76–87. doi:10.1037/0021-9010.81.1.76

Warr, P., Bartram, D., & Martin, T. (2005). Personality and sales performance: Situational variation and interactions between traits. *International Journal of Selection and Assessment, 13*, 87–91. doi:10.1111/j.0965-075X.2005.00302.x

Warren, R. A. (2008). *LMAP: Methods and statistical summary*. Retrieved from http://www.lmapinc.com/uploads/LMAP_Methods_&_StatisticalSummary.pdf

Waters, L. K. (1965). A note on the "fakability" of forced-choice scales. *Personnel Psychology, 18*, 187–191. doi:10.1111/j.1744-6570.1965.tb00277.x

White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 291–295. doi:10.1111/j.1754-9434.2008.00049.x

Wolford, K., & Christiansen, N. D. (2008, April). Effects of self-coaching on faking of personality tests. In R. L. Griffith & M. H. Peterson (Chairs), *Complex*

problems, simple solutions: Contemporary research in applicant faking behavior. Paper presented at the 23rd annual conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Wood, D., & Roberts, B. W. (2006). Cross-sectional and longitudinal tests of the personality and role identity structural model (PRISM). *Journal of Personality, 74*, 779–810. doi:10.1111/j.1467-6494.2006.00392.x

Zhao, H., & Seibert, S. E. (2006). The Big Five personality dimensions and entrepreneurial status: A meta-analytical review. *Journal of Applied Psychology, 91*, 259–271. doi:10.1037/0021-9010.91.2.259

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*, 551–563. doi:10.1037/0021-9010.84.4.551

Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology, 61*, 309–348. doi:10.1111/j.1744-6570.2008.00115.x

CHAPTER 29

WORK SAMPLE TESTS

George C. Thornton III and Uma Kedharnath

Work sample tests (WSTs) are high-fidelity assessment techniques that present conditions that are highly similar to essential challenges and situations on an actual job. WSTs have been used for a variety of purposes such as selection, certification, training, and performance evaluation for a wide variety of jobs.

Two quintessential work sample tests follow.

1. Applicants for a maintenance mechanic position in a large commercial construction company are required to complete a welding task in which they use an arc welder to build a steel frame to conform to a set of written specifications.
2. Candidates for certification as a software programmer are required to write computer code to apply discounts to orders of varying amounts of different products of the company.

Because WSTs are defined somewhat differently in different sources, we begin with definitions of terms as we use them here and some implications. We then describe a continuum of fidelity to the job and compare WSTs with assessment techniques with higher and lower levels of fidelity. Next, we provide examples of how work samples have been used for various purposes in work organizations and education for jobs with varying complexity. Scrutiny of these examples highlights the need to explicate several aspects of fidelity and points out the trade-offs of building higher versus lower levels of fidelity into a WST. Next, we make the often overlooked distinction between the method of the WST and the constructs that a WST measures. We point out that most literature does not specify what a WST measures. That discussion leads to a summary of evidence related to validity of work samples. The foundation for psychometric quality begins with the test construction process, and thus we provide three frameworks for constructing WSTs that ensure that several aspects of fidelity are built into any WST. Sprinkled throughout the chapter are many examples. We end the chapter with a discussion of the research needed to fill gaps in what is not known about WSTs and then offer conclusions.

DEFINITIONS OF WORK SAMPLE TESTS AND FIDELITY WITH IMPLICATIONS

Basic definitions of WSTs were first provided by Guion in 1965—"Work samples . . . sample directly the kind of behavior required by the job" (p. 195)—and then later in 1998—"a standard sample of a job content domain taken under standard conditions" (p. 509). We provide a more detailed definition of WSTs that explicates several dimensions of the notion of fidelity to the job.

A WST is a standardized and complex set of stimulus materials that has a high level of similarity and fidelity to a critical portion of a job presented to an examinee who is required to produce a product or demonstrate complex observable behavioral responses representative of critical job-relevant knowledge, skills, or abilities. Our definition is very similar to that of Ployhart, Schneider, and Schmitt (2006) and Roth, Bobko, and McFarland (2005), but we expand the definition with an elaboration of various dimensions of fidelity. Our definition is different from that of

DOI: 10.1037/14047-029
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

Asher and Sciarrino (1974) because we exclude tests of written knowledge and situational judgment and that of Truxillo, Donahue, and Kuang (2004) because we distinguish among exercises used in assessment centers (ACs). We show how some WSTs meet our definition and some do not. We weigh in on the discussion of how to determine criticality and how large a portion of a job should be covered by a WST.

Dimensions of Fidelity

Several dimensions in our definition of fidelity beg for expansion.

- *Physical fidelity:* The stimuli presented to the examinee entail complex testing materials, not just paper-and-pencil forms or questions on a computer screen. This procedure follows Truxillo et al. (2004). For example, applicants for a bank teller position could be asked to examine checks, verify signatures and account numbers, and dole out bills and coins.
- *Content fidelity:* The substance of the problems and challenges in the materials are highly similar to those of an actual job. For example, asking candidates for certification as human resource managers to discuss advantages and disadvantages of various employee benefit plans and to recommend the most desirable has high content fidelity, but asking them to discuss and choose among objects in the desert survival exercise (D. W. Johnson & Johnson, 1994) does not.
- *Situational fidelity:* The context of the WSTs is highly similar to the actual work setting and organization. For example, in a WST for customer service representatives applying to a call center, a high level of time pressure matching the job can assess whether candidates respond quickly, clearly, and accurately to inquiries.
- *Behavior fidelity:* The examinee must demonstrate a complex set of overt behavioral responses as he or she would on the job, not just select among proffered alternatives, make key strokes, or state behavioral intentions. For example, a teacher certification WST can require candidates to explain a complex topic in their area of expertise and answer questions from role player acting as a student.
- *Psychological fidelity:* The responses are indicators of attributes, such as knowledge, skills, and abilities, related to job performance (Goldstein, Zedeck, & Schneider, 1993). This feature is not unique to WSTs; in fact, psychological fidelity is the sine qua non of all predictors.

As with any test, a WST is only a sample; here, the sample is an important sample of the domain of tasks in the target job. This sample may be relatively narrow, but it is essential for effective job performance. In addition, as with all tests, the administration and scoring of WSTs is controlled; here, the scoring of behavioral responses or work products is standardized. These features distinguish WSTs from other high-fidelity assessments that are sometimes carried out on the job itself and may differ from one candidate to another.

Implications of Dimensions of Fidelity

The dimensions of fidelity of WSTs have strong implications for their design, administration, scoring, and evaluation. In this section, we provide examples of how high fidelity on each dimension might look in a specific WST and discuss the implications. One implication is that different methods of job analysis may be particularly helpful when establishing fidelity on different dimensions. In the following paragraphs, we suggest analytic methods described in Brannik and Levine (2001) and Perlman and Sanchez (2010).

The physical fidelity of WSTs comes from different aspects of the complex stimulus materials: complex passages to read, complicated and possibly vague instructions, live or video tapes of human beings, or physical objects, tools, and equipment. The materials may have varying levels of physical fidelity. For example, a high-fidelity WST for a manufacturing assembly job calls for tools, materials, and work space like those on the job. Any job analysis method that involves direct observation of actual job performance on the work site will yield information about the complexity of materials to build into a high-fidelity WST.

Fidelity in the content means the WST contains the problems and challenges encountered on the job. For example, the content of a WST to assess the

ability of marketing managers may require candidates to prepare a written sales plan for a specific product for a specific region. This feature implies the designer has detailed information about the most critical tasks on the job to build into the WST. Criticality is typically determined by ratings of frequency and importance of job tasks. Methods of work analysis, such as functional job analysis (Brannik & Levine, 2001; Perlman & Sanchez, 2010), that are applicable to a wide range of jobs and task inventories applicable to each specific job would be particularly helpful. Criticality could also include an assessment of hazards and how essential it is for each incumbent to complete the task.

Situational fidelity ensures that features such as the industry and setting, time pressure, and stress match the actual conditions on the job. For example, a one-on-one role-play activity may call for a candidate for promotion to captain in a police department to serve as public information officer and answer news reporters' rapid-fire questions. Situational analysis (Thornton & Mueller-Hanson, 2004) identifies contextual features of the job as the appropriate setting for the WST. One issue the organization faces is whether those conditions may cause adverse applicant reactions and whether they may not be an essential feature of the work of all police captains. Another issue is that the role players may not put the same level of stress on all candidates, thus raising the allegation of inconsistent and unstandardized administration.

WSTs have behavioral fidelity, that is, examinees must display complex overt behavior such as on-the-job behavior. This means a WST is a performance test (Anastasi & Urbina, 1997): Observers can see the process that the examinee goes through in dealing with the challenge of the WST. Worker analysis methods such as the Position Analysis Questionnaire provide a comprehensive framework to study behaviors in a wide variety of jobs, whereas time and motion studies focus on physical behaviors. High fidelity in an exam at the end of a training program for financial analysts may ask trainees to sift through a variety of data, compute return on investment, present recommendations and rationale to colleagues, come to consensus, and answer challenging questions from assessors playing the role of an executive committee. Thus, raters must observe overt complex behavioral responses and make judgments of the quality of responses and products. An issue is that it may be difficult for raters actually to observe the behaviors of multiple participants in the fast-moving and disperse actions of complex interactions. To reduce subjectivity and to make judgments more objective, various scoring aids such as rubrics are often used. Such steps to reduce subjectivity may constrain assessors to accept only a limited number of behaviors that could effectively handle the work sample.

Psychological fidelity implies that the WST is measuring the knowledge, skills, abilities, and other characteristics required for effective job performance. Although many job analysis techniques yield information about attributes needed for effective job performance, the job element method directly evaluates these human attributes. This aspect raises the issue of what constructs the WST is actually measuring and implies that the WST is neither deficient in measuring essential constructs nor contaminated with irrelevant variables. For example, reading ability may contaminate a physical ability test for truck drivers that presents complicated written instructions that are not representative of materials the job incumbent must read. We come back to this issue in our discussion of the validity of WSTs.

Our definition of WST includes two common requirements of a good test: a good sample and standardization. To say that a WST is a sample immediately raises the question of whether the sample is an adequate sample, that is, is it a large enough sample and is it representative of various facets of a domain? For example, a welder may be asked to use a stick welder with steel pieces, but the entire job may involve using aluminum and copper pieces with tungsten inert gas and metal inert gas welding equipment. The issue is whether the testing equipment is representative of a broad enough sample of the equipment used in the actual job. A WST need not cover all tasks of a job or tasks that new employees are expected to perform immediately on entry into the job; it should cover all new tasks that new employees are expected to perform.

Standardization implies consistency in administration and scoring. Standardization in the administration of work samples that call for interactions

with other people such as role players, assessors as challengers, and other candidates may be problematic. Various ways that role players interact differently with different candidates must be considered; some assessors may be harsher with follow-up questions, and the composition of participants in group discussions and games may affect the individual's opportunity to show his or her competencies. The issue is whether precise standardization across candidates is essential. It may very well be that some variation in interactions with others may provide a more thorough and accurate assessment.

CONTINUUM OF FIDELITY OF ASSESSMENT TECHNIQUES

In this section, we briefly describe several related methods with a wide range of fidelity between the test and the performance domain (see Figure 29.1). For most jobs, the highest fidelity assessment is job probation and the lowest fidelity assessments are paper-and-pencil instruments. A paper-and-pencil WST may possess high fidelity for those occupations that involve many paper-and-pencil tasks, for example, teachers and administrators. High-fidelity measures are quite concrete, whereas low-fidelity tests are abstractions of the job. High-fidelity tests involve samples of job behavior, whereas low-fidelity tests involve signs of behavior (Wernimont & Campbell, 1968). The simulations that fall in the middle of this continuum have different combinations of high- and low-fidelity indicators. In this chapter, we focus on a narrow portion of the measures near the top of this continuum: We cover true work samples and very high-fidelity simulations that are sometimes included in such processes as the AC method.

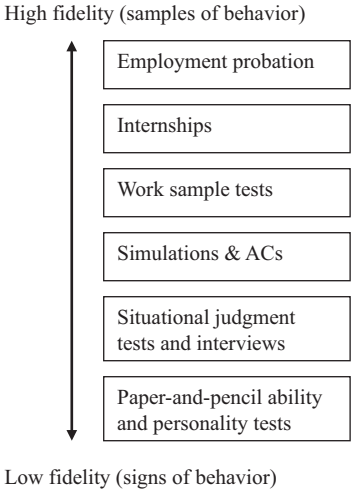

FIGURE 29.1. Continuum of fidelity of assessment methods. ACs = assessment centers.

Both signs and samples may be effective predictors of job performance; the challenge is deciding when to use either type of test. Low-fidelity signs may be appropriate when people entering a field must have a general aptitude for learning but will receive extensive training after selection. For example, in the medical field, measures of general aptitude may be appropriate for selection to medical school. By contrast, high-fidelity work samples may be more appropriate for highly specialized jobs in which the person must be highly skilled immediately on entering the job. For example, in the medical setting, a high-fidelity performance sample of a cardiologist would be appropriate.

Although the dimensions in our definition of fidelity clarify our meaning of work samples, not all distinctions are complete. It is still somewhat difficult to distinguish a work sample from other related types of measures. Anastasi and Urbina (1997) expressed this frustration: "Work samples and simulations merge imperceptibly" (p. 492).

Continuum of Fidelity Among Techniques

Examples of the lowest fidelity tests are multiple-choice cognitive ability tests and self-report personality questionnaires. We consider multiple-choice and self-report tests to be the lowest fidelity simulation because they appear to be quite different from the actual job (i.e., they lack face validity) and tend to measure abstract characteristics assumed to be related to job performance. These tests are called *signs* because they are signals of observable behaviors, not observable behaviors themselves (Thornton & Mueller-Hanson, 2004).

The next level of tests up the fidelity continuum includes situational judgment measures (Chan & Schmitt, 2005; Whetzel & McDaniel, 2009). These methods have been used for selection purposes (e.g., McDaniel, Hartman, Whetzel, & Grubb, 2007; Motowidlo, Dunnette, & Carter, 1990; Weekley &

Jones, 1999). They entail job-related scenarios that describe a problem or dilemma. Respondents are expected to apply their knowledge, skills, abilities, or other characteristics to respond to the scenarios (e.g., Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001). Situational judgment tests are administered in paper-and-pencil format or electronically. The respondent is usually given various response options from which he or she can choose the most appropriate response for each scenario. Situational judgment questions in the interview setting typically include hypothetical questions about how the interviewee would handle a situation (e.g., "What would you do if you were to get an angry customer and the customer service line was already long?"). Tests and interviews of situational judgment focus on job applicants' intentions to apply their knowledge, skills, abilities, or other characteristics to respond to the scenarios but do not require the applicant to actually demonstrate behaviors required on the job. Chapter 30 in this volume includes a full account of situational judgment measures.

The next level on the fidelity continuum consists of simulations and ACs (Thornton & Rupp, 2006). Simulation exercises are traditionally situational tests that use standardized procedures to elicit a sample of work-related behavior. The situations are very much like the actual job but are not replicas. The content includes the types of problems encountered on the job, but they may not be the exact, current organizational problems. Simulations call for observable and specific behaviors or skills that are needed for the job (Thornton & Mueller-Hanson, 2004). The exercises in ACs are more complex than the typically simple statement of problems and responses that express behavioral intention in situational judgment tests and interviews (for more information on these measures, see Chapter 30, this volume). ACs are considered to be higher fidelity than situational judgment tests and interviews because the assessees must exhibit complex, overt, and observable behaviors in the process of solving the hypothetical dilemmas, in contrast to statements of behavioral intentions in situational judgment measures. Trained assessors observe the assessees' overt behaviors and rate the assessees' performance on dimensions that are considered important for certain jobs, such as leadership skills for a managerial position (Thornton & Gibbons, 2009).

All exercises in ACs are not work samples, but some may be. AC exercises may have content that is different from the actual job; for example, exercises in a large financial organization may depict a situation in which assessees are required to solve a particular type of problem to perform successfully in the WST, when the target job actually requires employees to solve different types of problems. However, one large South African company posed current organizational issues in the leaderless group discussion exercise, making the exercise a work sample because it replicated a real task in the job of middle managers in that organization.

WSTs make up the next level on the continuum. WSTs are considered to be higher fidelity than ACs because they do not use hypothetical situations. Rather, they require that a candidate provide a sample of work-related behavior by completing a portion of the tasks that he or she would perform for the actual job (Roth et al., 2005). The behaviors assessed in work samples are very similar to those assessed in ACs in that they are observable and specific. However, the behaviors assessed in WSTs tend to be more complex and job specific than the behaviors observed in most AC exercises.

The level of fidelity above WSTs is internships. *Internships* are defined as structured and career-relevant work experiences that students obtain while they are still attending school (Taylor, 1988). Good internships are higher fidelity than work samples because interns tend to complete more than a small portion of the tasks that are required to perform well at the particular job. Unfortunately, not all internships involve important tasks; a carpentry intern may do little more than a few lowly activities such as cleaning up the job site.

Finally, the type of assessment with the highest fidelity is employment probation. *Employment probation* is a selection procedure in which new employees are hired under the condition that they will be retained after a specified probation period if their performance meets or exceeds expectations and terminated if their performance does not (e.g., De Corte, 1994; Ichino & Riphahn, 2005). We consider employment probation to be the highest

fidelity assessment because the organization sees a full range of employees' performance in real job settings.

Comparison of Work Sample Tests With Other Methods

A main factor that differentiates WSTs from moderate- to low-fidelity assessment tools is that WSTs call for demonstration of overt, specific, and complex behavior, production of a complex product that is identical to what is produced on the job, or both. Moderate- to low-fidelity measurement tools may test applicants' reasoning abilities, personality traits, and intelligence as predictors of job success, but these measurement tools do not call for demonstrations of actual behavior required on the job. Work samples are also useful for observing assessees' interactive skills that cannot be measured with low-fidelity instruments. For example, WSTs can be used to observe an insurance salesperson's interactions with potential clients or a medical intern's interactions with patients (Howard, 1983; Rupp & Searles, 2011).

Methods at the top of the continuum have more fidelity on more facets, but WSTs are more standardized than internships and employment probations. Whereas internship and job probation experiences tend to vary within and between organizations, all candidates participate in the same activities in WSTs. In addition, the scoring involved in work samples is more systematic than the performance evaluations in most internships and work probations.

As a general rule, the cost of assessment methods is greater at higher levels of fidelity. Whereas paper-and-pencil tests may be relatively inexpensive to administer and score on line, interviews may be more expensive because of the time to prepare, conduct, and evaluate the interview. Although some off-the-shelf simulations cost relatively little, the cost of administration, observation, and scoring can be considerable. The true cost of an AC includes a full accounting of development time, assessors, and expenses for lodging, meals, and rooms if assessors come from remote locations. The cost of work samples can be comparable to ACs. Hiring a person on probation incurs the cost of salary and benefits for the new employee plus the cost for training and supervision.

Although we primarily discuss high-fidelity measures in this chapter, we want to emphasize that we do not endorse one type of measurement tool over the others. Each type of measurement tool can make a contribution to assessment for selection and training. We do, however, endorse incorporating multiple measurement tools and techniques when possible to obtain different kinds of information and therefore assess a bigger part of the picture.

USES OF WORK SAMPLE TESTS

WSTs have been used for several different purposes in jobs with differing levels of complexity. Some of the applications of WSTs are described next.

Selection

WSTs are most often used in personnel selection (e.g., Callinan & Robertson, 2000). They have appeal for selection because they are among the most valid predictors of job performance (Hunter & Hunter, 1984; Jackson, Harris, Ashton, McCarthy, & Tremblay, 2000; Schmidt & Hunter, 1998). They have relatively low levels of adverse impact and subgroup differences (Callinan & Robertson, 2000), but recent analyses have shown they may not be as immune to racial differences as once thought (Dean, Roth, & Bobko, 2008). Speculation has abounded about the reasons for any racial differences and any changes in differences over time: differences in comparison samples in earlier and later studies, increasing real differences in achievement gaps, opportunities to learn occupationally specific skills or to acquire the necessary knowledge, and so forth. Research is needed to study these knotty issues. Aside from differences in group means, it would be important to know whether the use of WSTs in high-stakes selection results in prediction bias. To our knowledge, research on the issue has not been reported.

Research has shown, however, that applicants' reactions to work samples have generally been positive (Hausknecht, Day, & Thomas, 2004). Applicants react positively to selection procedures that exhibit a strong relationship to the job content, appear to be fair, are not administered in a paper-and-pencil format, and appear to have a face-valid

format (Rynes, 1993; Rynes & Miller, 1983), and WSTs meet these criteria.

The two types of selection scenarios are achievement oriented and aptitude oriented. In the former, the candidate is expected to be able to perform the job immediately on selection; here, all aspects of fidelity are high. In the aptitude scenario, the candidate must have the basic abilities to do the job but will receive some training on organization-specific equipment or processes; here, behavioral and psychological fidelity will be high, but physical, content, and situational fidelity may be somewhat lower.

Robertson and Downs (1989) described the difference between normal WSTs and trainability tests: Normal WSTs are used for people who are already trained, and trainability tests are used for people who are not trained and when a learning period is expected. For more information on trainability tests, see Robertson and Mindel (1980). Robertson and Downs conducted a meta-analysis on WSTs of trainability and found that trainability WSTs tend to predict short-term training success more accurately than longer term training success. They also suggested that greater situational variability may occur when short-term follow-up periods are used.

Certification

WSTs are used for the certification of professional competence in a variety of jobs. For example, the teacher work sample methodology of Western Oregon University (Schalock, 1998) assesses preservice and in-service teachers' standing on national and state teaching standards and their impact on their students' learning (Denner, Salzman, & Bangert, 2001). The work samples collected from teachers help to differentiate performance along the continuum from beginning to expert teaching. Denner et al. (2001) examined the validity and reliability of this test and found some evidence that it can credibly connect teachers' performance to learning.

Work samples are used in many education settings, often with the goal of assessing competency in teachers or for selection purposes. An applied example of a work sample in the educational setting is a teacher work sample used to assess preservice teachers' use of technology in their future career (Graham, Tripp, & Wentworth, 2009). Preservice teachers are typically college students who get guidance and supervision from currently employed teachers, who act as their mentors. The Teacher Work Sample (Renaissance Partnership, 2001) requires preservice teachers to develop and implement a teaching plan during their field experience, and it is meant to assess the candidate's best work. In this case, the raters are groups of faculty members. See Chapter 20 in this volume for an account of performance tests in education.

Another way in which WSTs can certify professional competence is to set cutoff scores and standards against which applicants' scores can be compared. Cascio and Aguinis (2005) described using work samples to set minimum standards and define low and high proficiency in the analytical judgment method. Work samples for each question are given to panelists who rate each examinee's work sample on a classification scale, for example, starting from basic and working up to proficient and advanced. The average score on the work samples becomes the point estimate of that performance standard. A full account of credentialing exams is presented in Volume 3, Chapter 19, this handbook.

Skill-Based Pay

Using work samples to certify professional competence ties into using work samples to determine skill-based pay for employees. Skill-based pay is distinct from traditional job-based pay because pay is contingent on employees acquiring and demonstrating proficiency in a new skill (Shaw, Gupta, Mitra, & Ledford, 2005). For example, firefighters may be given skill-based pay for demonstrating proficiency in the use of a new piece of first aid equipment.

Training

Work samples can be used to evaluate the effectiveness of training programs. For example, if the objective of a training program is to teach employees how to use a specific software program, then a WST can be administered at the end of the training to determine whether the training objectives were achieved. The effectiveness of training can also be evaluated by giving WSTs to job incumbents who have been working for several months after the training. If the incumbents are able to do the tasks at a predetermined

level of competence, then the organization can assume that the knowledge and skills learned in training have been transferred to the work setting (Felker, Curtin, & Rose, 2007).

Criterion Measurement

WSTs can also be used as criteria for proficiency in job performance. Criterion measures are used to determine pay, job retention, and performance levels in validation of selection methods. Some of the exemplary work in this use of WSTs can be seen in the military. Carey (1991) used measures of job performance to validate training outcomes for U.S. Marine Corps duties such as assembling radios, preparing a launcher for firing, or throwing dummy grenades. Hedge and Teachout (1992) examined the feasibility of using interviews to measure work sample criteria. They administered hands-on tests and interview WSTs to Air Force personnel across various job specialties. In the interview approach, candidates were asked to describe how they would perform a task. The candidates were assessed with a method similar to hands-on testing. Hedge and Teachout concluded that the two types of assessment methods were equivalent and resulted in the same rank ordering of candidates. They also concluded that the two assessment methods were not equivalent for diagnosing training needs because the correlations at the task level between the two methods varied between and within job specialties or areas.

Campbell and Knapp (2001) described extensive use of WSTs as one set of criteria of job performance in Project A for the U.S. Army. WSTs were developed to measure job performance for nine select military occupations and supervisors at the end of training during the first 3-year tour of duty and during the second tour of duty during Years 4 and 5. Examples include repairing a vehicle, safety and survival tasks, and personnel counseling with a role player. These and other measures were used as criteria against which experimental and existing predictors of job performance were validated.

WORK SAMPLE TESTS FOR JOBS OF DIFFERENT COMPLEXITY: O*NET

The term *work sample* begs the question: a sample of what? The purpose of this section is to describe different job levels at which WSTs can sample performance-related behaviors. We use the Occupational Information Network (O*NET) database as a framework to describe these levels because O*NET provides an elaborate structure to describe the different aspects of jobs. O*NET is an ongoing research program sponsored by the U.S. Department of Labor (Peterson & Sager, 2010). It classifies a large number of occupations and replaced the 70-year-old *Dictionary of Occupational Titles* (U.S. Department of Labor, 1991). O*NET is available online and is updated periodically. A major benefit of using the O*NET database is that a common language is used to categorize occupations in terms of six domains: (a) worker characteristics, (b) occupation characteristics, (c) worker requirements, (d) occupational requirements, (e) occupation-specific requirements, and (f) experience requirements. These domains include a general categorization framework under which more specific occupational information can be organized. For example, the worker characteristics domain includes information regarding the required abilities, interests, work values, and work styles for particular jobs (Peterson et al., 2001).

The use of a common language allows users to describe and contrast occupations with relative ease (LaPolice, Carter, & Johnson, 2008). O*NET has been used for several purposes, including providing information to develop assessment systems to select and promote employees (Peterson et al., 2001).

On the basis of the level of experience, education, and training that an individual needs to do a job, O*NET provides a taxonomy of job zones, with Zone 1 jobs requiring the least preparation and Zone 5 jobs requiring the most preparation. Examples of work samples at all levels of jobs are presented next.

Zone 1 jobs need little or no preparation, including fast-food cooks, parking lot attendants, and construction laborers. An organization interested in using work samples to hire construction laborers might ask applicants to position, align, and seal pipes; unload and identify building materials; and position forms for pouring concrete.

Jobs in Zone 2 require some preparation. This zone includes jobs such as animal trainers, bus drivers, and receptionists. If one is interested in selecting bus drivers for a public transit job, one might

ask applicants to drive vehicles on specified routes according to the time schedule and according to traffic regulations, park the vehicle at pickup areas for passengers, and load and unload bags in baggage compartments.

Jobs in Zone 3 require medium preparation and include jobs such as air traffic controllers, acute care nurses, electricians, and plumbers. A WST for plumbers might include measuring, cutting, threading, and bending pipe to specified angles and using the appropriate power tools (e.g., pipe cutters) to complete the task. Other tasks may include studying building plans and inspecting structures to assess the materials and equipment needed, establishing the sequence of pipe installations, and planning the installation around obstacles such as wiring.

Jobs in Zone 4 need considerable preparation and include accountants, personnel recruiters, and computer programmers. A WST for computer programmers may include correcting errors and rechecking the program, writing and analyzing programs, and performing systems analysis. Jobs in Zone 5 need extensive preparation and include biologists, dentists, and nurse practitioners. A WST for nurse practitioners may include prescribing medication dosages and frequencies on the basis of patients' characteristics, developing treatment plans on the basis of scientific rationale and professional practice guidelines, and prescribing medications on the basis of safety and cost.

It is important to note that the work samples that are chosen to be included in a selection test are generally considered to be the most important or essential tasks of the job. Systematic task analysis is recommended to decide which tasks are the most central for a job in each organization.

EVIDENCE OF RELIABILITY AND VALIDITY

In this section, we review the types of evidence of reliability and validity that have been marshaled in the past to support various inferences from scores on WSTs. First, we cover research findings on the reliability of WSTs. Then, we discuss the various types of evidence that have been found for the validity of WSTs. We endorse the basic principle of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) that conclusions about the validity of test scores should be made on the totality of accumulated evidence about the use of test scores (see Chapter 13, this volume, for a full discussion of the *Standards*). Note that some terms for types of evidence, such as *content, criterion,* and *construct validity,* are now outdated. We summarize this section with conclusions about the overall validity of WSTs.

Reliability

Roth et al. (2005) found four test–retest reliability estimates for work samples (.76, .74, .71, and .61) and used these values to correct for attenuation in the WSTs. A reliability coefficient of .60 has been found in previous literature for supervisory ratings of performance (e.g., Viswesvaran, Ones, & Schmidt, 1996), and a reliability coefficient of .80 has been found in previous literature for objective performance measures (e.g., Roth, Huffcutt, & Bobko, 2003). Test–retest reliability for WSTs has been supported in the context of job performance in the military. For example, Mayberry (1990) studied infantrymen in the Marine Corps and found a test–retest reliability of .77 with an alternate form of a work sample test. Test–retest reliabilities for automotive mechanics and helicopter mechanics were reported to be .79 and .88, respectively, and split-half reliability estimates for automotive mechanics and helicopter mechanics were .92 and .97, respectively (Mayberry, 1992).

In addition to evidence of test–retest reliability, some evidence of high interrater agreement on scores for WSTs has been found, especially in the military context. For example, Felker et al. (1988) and Carey (1990) found interrater agreement higher than .90, and other researchers have found pairwise agreements between .74 and .90 across different teams and occupations (Doyle & Campbell, 1990; Hedge, Lipscomb, & Teachout, 1988). More evidence along these lines is needed, as Roth et al. (2005) said in their meta-analysis that they could not find many studies that reported reliabilities. Variation in reliability estimates, the range of existing estimates, and the dearth of studies with

comparable data for paper-and-pencil tests may be due to the time demands of administering and scoring WSTs and the inherent difficulties of developing totally objective scoring methods. Caution is also warranted in generalizing from the results of meta-analyses of reliabilities of WSTs because studies have examined many different types of WSTs.

Content Representativeness (Content Validity)

In the context of WSTs, content representativeness addresses whether the content of the test is shared with the content of the actual job. In other words, a test is said to have content evidence of validity if the content of and the performance needed to do well on the test is highly similar to the content of and the performance needed to do well in the actual job. This type of evidence is strong if the test is based on a systematic job analysis including information on working conditions, tools and materials used, and job requirements (Felker et al., 2007). For example, Denner et al. (2001) looked for and found initial evidence that the teacher work sample tasks they were using aligned with the national, state, and institutional standards (e.g., planning for and teaching actual lecture material).

Correlations With Performance Criteria (Criterion, Concurrent, or Predictive Validity)

A considerable amount of research has been done on the correlation of WSTs with measures of work performance. The criterion may be a score or rating that is available at the time the predictor is measured (i.e., concurrent validity) or later (i.e., predictive validity). Hunter (1983) performed a meta-analysis of WSTs and reported a work sample supervisory rating mean correlation of .42 ($K = 7$, $N = 1{,}790$) for nonmilitary studies and a mean correlation of .27 ($K = 4$, $N = 1{,}474$) for military studies, correcting for criterion unreliability.

A commonly cited source for the validity of work samples is the meta-analysis conducted by Hunter and Hunter (1984). They reported a validity coefficient of .54 for WSTs predicting supervisory ratings, correcting for the unreliability of the supervisor ratings. Schmitt, Gooding, Noe, and Kirsh (1984) conducted a meta-analysis including studies published between 1964 and 1984 in the *Journal of Applied Psychology* and *Personnel Psychology* and found an uncorrected validity coefficient of .32 ($K = 7$, $N = 382$) when the criterion was ratings on job performance. Russell and Dean (1994, as cited in Roth et al., 2005) extended Schmitt et al.'s (1984) study by focusing on research published between 1984 and 1992. They reported a validity coefficient of .37 ($K = 20$, $N = 3{,}894$) in relation to job performance across a variety of jobs. Using a criterion of trainability, Robertson and Downs (1989) found the correlation for WSTs was significant, albeit lower than that for cognitive ability tests (.41 vs .56, respectively).

More recently, studies have found lower validity estimates than those previously obtained. For example, Roth et al. (2005) reported a corrected validity coefficient of .33, which is noticeably lower than the estimates reported by Hunter and Hunter (1984) and Schmidt and Hunter (1998), that is, .54. They concluded that research could overestimate work sample validity. Although some recent validity estimates have been lower than those previously obtained, these newer estimates took into account some important limitations that have been noted in previous literature (e.g., conceptual and methodological limitations; see Asher & Sciarrino, 1974). Differences in meta-analytical estimates of validity may be due to the sampling of jobs, examinees, type of WST, and research reports aggregated. Also, different corrections for range restriction and unreliability of measures have been used. Overall, recent studies have found lower validity coefficients for WSTs but still support the idea that WSTs are valid predictors of job performance.

Incremental Validity

The results found for the incremental validity of WSTs have been mixed. For example, Schmidt and Hunter (1998) found that adding a WST to a general mental ability or cognitive ability test would add .12, or a sizable 24% increase in correlation. Roth et al. (2005) found an incremental correlation of .06 of WSTs over cognitive ability tests alone, that is, a 12% increase. Overall, studies have shown that WSTs have incremental validity over cognitive ability tests.

Demographic Differences

WSTs have been known to show smaller subgroup differences than other selection tools such as cognitive ability tests. For example, previous research has consistently found a subgroup effect size of .38 for Black and White comparisons in WSTs and no subgroup differences when comparing Hispanics and Whites (Schmitt, Clause, & Pulakos, 1996). Schmitt et al. (1996) also found that WSTs favored women slightly over men ($d = 0.38$). Schmitt and Mills (2001) found smaller subgroup differences on simulations than on paper-and-pencil tests. However, more recently, Roth, Bobko, McFarland, and Buster (2008) found a much higher Black–White difference ($d = 0.73$) among job applicants than had previous studies. Potential reasons for group differences were described earlier in this chapter.

Convergent and Discriminate Relationships (Construct Validity)

Salgado, Viswesvaran, and Ones (2002) noted that the research is very limited on the correlation of WSTs with other similar and different measures of similar and different constructs. Salgado et al. suggested that previous meta-analytic studies have found considerable generalizability across situations, which hints that different WSTs share a core construct. Among the limited literature in this vein is Schmidt and Hunter's (1992) process model, which predicted that one's performance on WSTs can be explained by the relationships of test performance to general mental ability, motivation, and experience.

Ones and Viswesvaran (1998) found correlations between WSTs and interviews (.20), overt integrity tests (.07), personality-based integrity tests (.27), and job knowledge tests (.36). Roth et al. (2005) found an overall correlation of .32 between WSTs and cognitive ability tests ($K = 43$, $N = 17{,}563$), which increased to .38 when correcting for work sample unreliability and increased to .48 when correcting for range restriction. They also reported a correlation of .13 between WSTs and situational judgment tests ($K = 3$, $N = 1{,}571$).

A relationship between physical ability tests and work samples that require significant muscular strength to complete tasks involving equipment, tools, or apparatus (e.g., using hand tools, attaching fire hose, carrying gear boxes) has been found in previous literature. Researchers have suggested that strength tests measure some of the same constructs that underlie these sorts of WSTs. This construct overlap may be the reason that strength measures have been strongly correlated with such work samples. The methodology and methods of measurement used to test the underlying constructs in strength tests and WSTs also appears to be similar; specific behaviors are tested in both (Blakley, Quinones, Crawford, & Jago, 1994).

Summary

These results begin to show what attributes WSTs measure. They measure constructs underlying performance on a variety of jobs. These constructs include something akin to general cognitive ability, specific abilities (e.g., problem solving), declarative and procedural knowledge, personality variables (e.g., conscientiousness and adaptiveness), and physical abilities. They are related to interviews, personality-based integrity tests, and job knowledge tests, but not to overt integrity tests.

THREE SETS OF GUIDELINES FOR BUILDING WORK SAMPLE TESTS

Whereas much has been written about building multiple-choice tests and self-report questionnaires, relatively little advice has been published about building WSTs. On the surface, the WST development process may seem simple: Just ask candidates to complete a task or two important to the job. In some cases that may be feasible; for example, some jobs entail only a very small number of highly repetitive narrow tasks. In most situations, however, jobs are more complex, and candidates cannot be asked to do the exact job itself because of physical, security, safety, or financial restrictions. In addition, even if candidates could do an actual job task, there are still many other considerations that make such a simple extraction rather unrealistic and unsatisfactory. Thus, the WST must be at least somewhat removed from the job itself. Furthermore, one also needs to consider time limits, instructions, administration, scoring, norms, or other ways to interpret the meaning of scores.

In this section, we provide three sets of suggestions for building a WST. First, we summarize steps described by Truxillo et al. (2004). Second, we expand guidance from a model for building simulations into ideas for building WSTs. Third, we show how considerations of the five aspects of fidelity help design WSTs for different purposes.

Steps

Truxillo et al. (2004) provided a detailed description of steps and practical considerations for the development, administration, and scoring of work samples and simulations. To prepare for the development of this process, designers must select the right subject matter experts. Subject matter experts must be willing to participate, have sufficient knowledge of the job, and possess good written and oral communication skills. The next step is to specify the performance domain. This specification might be done through traditional job analysis of tasks, critical incidents, or competencies (for a discussion of job analysis techniques, see Chapter 23, this volume, and Perlman & Sanchez, 2010). This step culminates in the specification of a defensible test plan. Developing the test entails preparation of task situations to present to examinees. If high-fidelity situations are to be used, the developer must compromise between replication of the job and practical considerations of time and expense. Practically, the designer must obtain high-quality information from subject matter experts and decide the format for presentation of test material, either on paper or via videos. Truxillo et al. also described issues in administering the test.

Model

Guidance for building a WST is provided by a model for building simulation exercises contained in Thornton and Mueller-Hanson (2004). Figure 29.2 shows that any test construction project begins with a thorough analysis of several aspects of the situation one wants to simulate. Central to building a work sample is the analysis of the tasks in the job in question. Any WST calls for the examinee to complete just one task or a very few tasks in the job. The task analysis must determine the most important, critical,

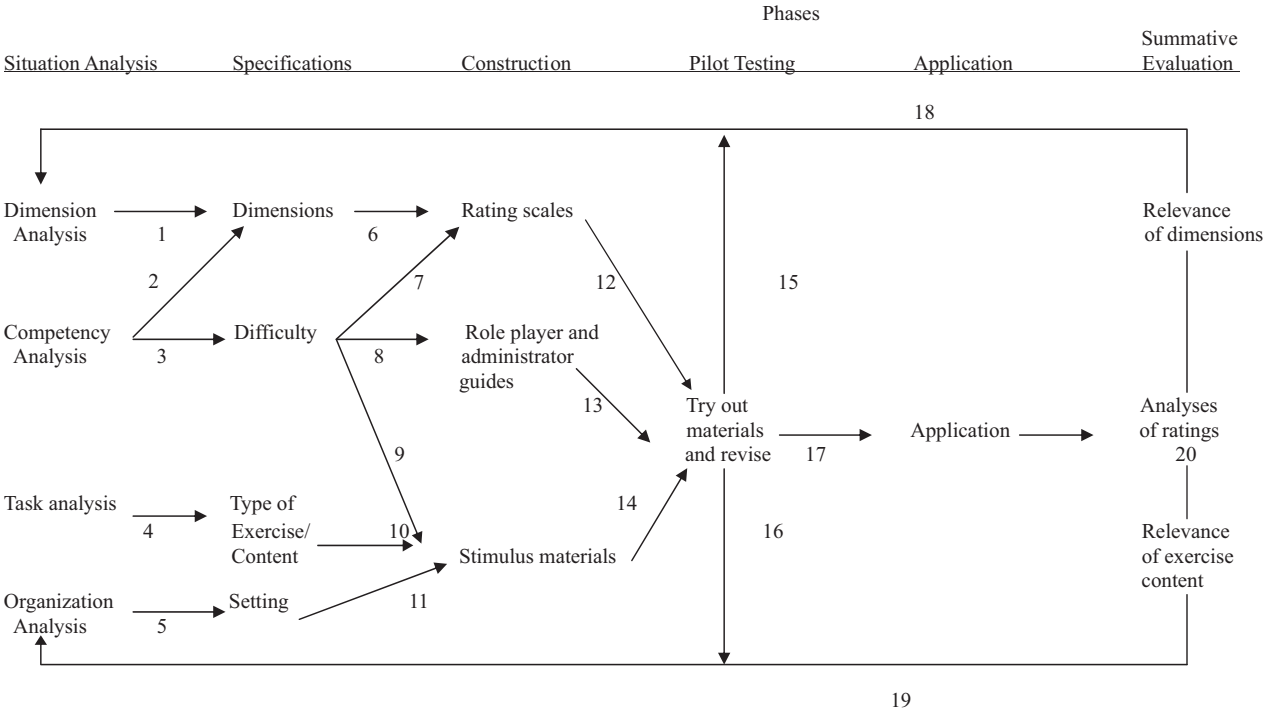

FIGURE 29.2. Model for constructing simulation exercises. Model 2.1 from *Developing Organizational Simulations: A How-To Guide for Practitioners and Students* (p. 18), by G. C. Thornton III and R. A. Mueller-Hanson, 2004, Mahwah, NJ: Erlbaum. Copyright 2004 by Lawrence Erlbaum Associates. Used with permission.

or essential task required of all job incumbents (Link 4 in Figure 29.2). If the work sample test is to be used for selection, the task must be one that all new employees must be able to accomplish when hired; it should not be a task new employees can be readily trained to do after selection. Possible accommodations for candidates with disabilities must be considered: If the job task is essential, then the organization is responsible for determining whether reasonable accommodations can be made so that individuals with various physical and mental disabilities can complete the task. If accommodations can be made in the job, then accommodations should be made in administering the work sample test.

The task analysis provides specifications for the content of the work sample test, including the instructions, equipment, supplies, work space, and so forth. On the basis of test specifications, actual testing materials can be selected or built (Link 10 in Figure 29.2). To illustrate this process, consider the construction of a work sample test for welders who must build a metal frame on the basis of a set of written instructions and drawings. The testing materials might include a welding machine, tools, various pieces of metal, safety equipment, instructions, and so forth.

A competency analysis provides guidance on how difficult the work sample will be (Link 3). Organizational members specify the level of proficiency that examinees must demonstrate to pass the test. Given the required level of proficiency, the assignment for the work sample task can be determined (Link 9): For the welding task, a simple or complex design can be required, and a loose or tight set of tolerances can be specified. The specified level of difficulty also provides guidance for constructing the rating scales and the level of accuracy that is needed to earn credit for task completion (Link 8). For example, on the welding work sample, the examiner can be told the required degree of accuracy in the angles of the final product or the time limits allowed for completing the task.

For some types of work samples, examinees may interact with the administrator or role players. The competency analysis will inform these individuals how to interact with the examinee, for example, whether to be supportive or obstructionist. For example, in the welding task, the administrator may entertain questions or provide no guidance or a role player may simulate the person commissioning the welded product and request a change in design or criticize the product so as to introduce work-related stress.

Finally, the situation analysis will inform the test developer of the dimensions of the behavior and output that will be scored (Link 1). The process of completing the work sample, the quality of the output, or both may be evaluated. On the welding work sample, the final welded piece may be evaluated on the accuracy of the angles, the smoothness of the joints, and so forth. Scrutiny of the welder's behavior may include observance of safety practices and maintenance of a clean and tidy workplace. Rating scales may be used to provide support to the examiner (Links 6 and 7). The rating scales may be simple Likert-type scales (e.g., poor to excellent) or more systematically developed behaviorally anchored rating scales. The latter are sometimes referred to as *rubrics*.

After all the testing materials are constructed, pilot testing is highly recommended before application. Summative and formative evaluation can then be undertaken. Further details on the specific steps of each phase of WST construction can be found in Thornton and Mueller-Hanson (2004).

Arguments for Less Than Maximum Fidelity of All Dimensions of Fidelity

There are compelling reasons for building WSTs with high fidelity, as noted throughout this chapter: They are among the most effective predictors of job performance, their face validity makes them appealing to candidates, and they provide measures of specific elements of job performance. All WSTs have a high degree of fidelity in comparison with other assessment techniques, as depicted in Figure 29.1. At the same time, there are also compelling arguments for not having the very highest levels of fidelity on all dimensions in all situations. When designing WSTs for certain purposes, some aspects of fidelity may be somewhat removed from the job. For example, consider the example of the WST for computer programmers to write software for figuring discounts. Performance may be contaminated by

insider knowledge of the exact specifications of the products and their advantages over other similar products. The issue the user faces is to decide how much he or she wants to assess this specific knowledge versus more general programming skills. To the extent that content and physical fidelity are heightened, the test may exclude people who have the basic skills but not the specific knowledge to score well. For some purposes, it may be ill advised for the WSTs to have the highest levels on all dimensions of fidelity.

Table 29.1 presents suggestions for the levels of five dimensions of fidelity desirable for WSTs used for several different purposes. The reader will quickly note that a very high level of psychological fidelity is suggested for all purposes. However, there may be advantages to reducing to some extent the level of fidelity on some dimensions for some assessment purposes.

WSTs might be used for selection of personnel who must be fully capable of doing the job from Day 1; we call this *achievement-oriented selection.* In such situations, the WST will have full fidelity in all dimensions. In a somewhat different scenario, the organization may want to select candidates with a strong aptitude to do the job, but newly hired staff will receive training in the specific equipment, methods, and so forth on the job. We call this *aptitude-oriented selection.* Thus, the WST may entail the use of a piece of equipment that is not on the job, or a situation that is not quite as dangerous as the most stressful actual job challenge. In general educational settings, in which students may end up pursuing different career paths, work samples may reasonably relax some fidelity requirements (R. L. Johnson, Penny, & Gordon, 2008). By contrast, in a technical training program for a specific occupation, work samples may have very high levels of fidelity on all dimensions.

RESEARCH NEEDS

More theory and research is needed to advance psychologists' understanding of the science and practice of the work sample method. In this section, we pose questions in five areas based on some of the needs mentioned throughout the chapter. First, what are the trade-offs in relaxing the highest level of one or more of the dimensions of fidelity on different outcomes such as predictive correlations? In other words, "What difference does it make when WSTs do not have maximum fidelity? Such studies would address Roth et al.'s (2005) call for future research on the fidelity of predictors.

Second, how broad of a sample of work behaviors in the job domain must be covered by the WST to justify making inferences about job competence? Third, we underscore Roth et al.'s (2005) call for continued focus on the constructs measured by WSTs. The literature on WSTs does not clearly make a distinction between method and construct (Arthur & Villado, 2008). It appears that WSTs often measure some ill-defined combination of declarative knowledge, procedural knowledge,

TABLE 29.1

Variations in Levels of High Fidelity for Different Applications

	Selection					
Aspects of fidelity	Achievement oriented	Aptitude oriented	General education	Specific job training	Criterion measurement	Certification
Physical fidelity	HH	H	H	HH	HH	H
Content fidelity	HH	H	H	HH	HH	HH
Situational fidelity	HH	H	H	HH	HH	H
Behavioral fidelity	HH	HH	H	HH	HH	H
Psychological fidelity	HH	HH	HH	HH	HH	HH

Note. HH = very high level of fidelity; nearly replicating the job; H = high level of fidelity; work sample clearly resembles the job.

general cognitive ability (*g,* intelligence), specific abilities (e.g., problem solving), various physical skills, and personality variables such as conscientiousness or adaptability. More research is needed to answer the question, "Just what psychological variables are measured by any given WST?"

Fourth, more generally, to what extent do work samples measure "can do" or "will do" variables? Stated differently, are WSTs maximum performance measures or typical performance measures? (See Klehe & Anderson, 2005, for a discussion of this distinction.) What kinds of contextual job performance and citizenship behaviors are predicted by WSTs?

Fifth, what legal issues arise in the use of WSTs? How have judges in employment discrimination litigation evaluated the job relatedness and business necessity of WSTs? The use of ACs has received mixed reactions in the legal arena (Thornton, Wilson, Johnson, & Rogers, 2009), but researchers need more information on these legal reactions. Is it legally defensible to measure only one narrow but essential job task? As noted previously in this chapter, research on the potential prediction bias in the use of WSTs for various subgroups covered by employment discrimination legislation is sorely needed. Finally, for any given WST what kinds of accommodations are needed to comply with the Americans With Disabilities Act? Discussion of many legal issues in testing is contained in Chapter 38 of this volume.

CONCLUSIONS

In this chapter, we made several theoretical and practical contributions to the understanding and application of WSTs. The theory of fidelity was extended by explicating five dimensions of fidelity, namely physical, content, situational, behavioral, and psychological. Furthermore, we questioned the assumption that high levels of fidelity on all dimensions are desirable and proposed that it may be preferable in some situations to somewhat reduce the requirement for high fidelity on some dimensions for some applications of WSTs. We note that relatively little theory or research has investigated the actual constructs measured by WST methods. Although considerable research has demonstrated that WSTs predict, with incremental accuracy, job performance criteria, more theory and research are needed to explain why this occurs. We propose that WSTs measure some combination of general cognitive ability, technical knowledge, and physical abilities. However, more research is needed to establish what facets of cognitive and physical abilities are operative in different WSTs.

This chapter included several practical suggestions. A wide variety of more than 30 examples of WSTs in prior studies and possible work samples might be used for jobs across the range of job complexity shown in O*NET job levels (e.g., welders, computer programmers). Three different guidelines for constructing WSTs were presented, including a model for building simulation exercises. Table 29.1 listed proposals for how the dimensions of high fidelity might be relaxed for different applications of work samples.

References

American Educational Research Association, American Psychological Association, & American Council on Measurement in Education. (1999). *Standards for educational and psychological tests* (3rd ed.). Washington, DC: American Psychological Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, NJ: Prentice-Hall.

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435–442. doi:10.1037/0021-9010.93.2.435

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27*, 519–533. doi:10.1111/j.1744-6570.1974.tb01173.x

Blakley, B. R., Quinones, M. A., Crawford, M. S., & Jago, I. A. (1994). The validity of isomorphic strength tests. *Personnel Psychology, 47*, 247–274. doi:10.1111/j.1744-6570.1994.tb01724.x

Brannik, M. T., & Levine, E. L. (2001). *Job analysis.* Thousand Oaks, CA: Sage.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8*, 248–260. doi:10.1111/1468-2389.00154

Campbell, J. P., & Knapp, D. J. (Eds.). (2001). *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Erlbaum.

Carey, N. B. (1990). *An assessment of surrogates for hands-on tests: Selection standards and training needs* (CRM 90–47). Alexandria, VA: Center for Naval Analyses.

Carey, N. B. (1991). Setting standards and diagnosing training needs with surrogate job performance measures. *Military Psychology, 3*, 135–150. doi:10.1207/s15327876mp0303_1

Cascio, W. F., & Aguinis, H. (2005). Test development and use: New twists on old questions. *Human Resource Management, 44*, 219–235. doi:10.1002/hrm.20068

Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *Handbook of personnel selection* (pp. 219–242). Marden, MA: Blackwell.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt-Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410–417. doi:10.1037/0021-9010.86.3.410

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93*, 685–691. doi:10.1037/0021-9010.93.3.685

De Corte, W. (1994). Utility analysis for the one-cohort selection-retention decision with a probationary period. *Journal of Applied Psychology, 79*, 402–411. doi:10.1037/0021-9010.79.3.402

Denner, P. R., Salzman, S. A., & Bangert, A. W. (2001). Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Education, 15*, 287–307. doi:10.1023/A:1015405715614

Doyle, E. L., & Campbell, R. C. (1990, November). *Navy: Hands-on and knowledge tests for the Navy radioman.* Paper presented at the 32nd annual conference of the Military Testing Association, Orange Beach, AL.

Felker, D. B., Crafts, J. L., Rose, A. M., Harnest, C. W., Edwards, D. S., & Bowler, E. C. (1988). *Developing job performance tests for the United States Marine Corps infantry occupational field* (AIR-47500–9/88-FR). Washington, DC: American Institutes for Research.

Felker, D. B., Curtin, P. J., & Rose, A. M. (2007). Tests of job performance. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 319–348). New York, NY: Taylor & Francis.

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of job analysis–content validity process. In N. Schmitt, W. C. Borman, & Associates. (Eds.), *Personnel selection in organizations* (pp. 3–34). San Francisco, CA: Jossey-Bass.

Graham, C. R., Tripp, T., & Wentworth, N. (2009). Assessing and improving technology integration skills for preservice teachers using the teacher work sample. *Journal of Educational Computing Research, 41*, 39–62. doi:10.2190/EC.41.1.b

Guion, R. M. (1965). *Personnel testing.* New York, NY: McGraw-Hill.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639–683. doi:10.1111/j.1744-6570.2004.00003.x

Hedge, J. W., Lipscomb, M. S., & Teachout, M. S. (1988). Work sample testing in the Air Force job performance measurement project. In M. S. Lipscomb & J. W. Hedge (Eds.), *Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel* (AFHRL-TP-87-58). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Training Systems Division.

Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion measurement. *Journal of Applied Psychology, 77*, 453–461. doi:10.1037/0021-9010.77.4.453

Howard, A. (1983). Work samples and simulations in competency evaluation. *Professional Psychology: Research and Practice, 14*, 780–796. doi:10.1037/0735-7028.14.6.780

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, and job performance, and supervisory ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257–266). Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98. doi:10.1037/0033-2909.96.1.72

Ichino, A., & Riphahn, R. T. (2005). The effect of employment protection on worker effort: Absenteeism during and after probation. *Journal of the European Economic Association, 3*, 120–143. doi:10.1162/1542476053295296

Jackson, D. N., Harris, W. G., Ashton, M. C., McCarthy, J. M., & Tremblay, P. F. (2000). How useful are work samples in validational studies? *International Journal of Selection and Assessment, 8*, 29–33. doi:10.1111/1468-2389.00129

Johnson, D. W., & Johnson, F. P. (1994). *Joining together: Group theory and group skills* (5th ed.). Boston, MA: Allyn & Bacon.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tests.* New York, NY: Guilford Press.

Klehe, U., & Anderson, N. (2005). The prediction of typical and maximum performance in employee selection. In A. Evers, N. Anderson, & O. Voskuijl (Eds.),

Handbook of personnel selection (pp. 331–353). Marden, MA: Blackwell.

LaPolice, C., Carter, G. W., & Johnson, J. W. (2008). Linking O*NET descriptors to occupational literacy requirements using job component validation. *Personnel Psychology, 61*, 405–441 doi:10.1111/j.1744-6570.2008.00118.x

Mayberry, P. W. (1990). *Validation of ASVAB against infantry job performance.* Alexandria, VA: Center for Naval Analyses.

Mayberry, P. W. (1992). Evaluating minimum aptitude standards. *Military Psychology, 4*, 1–16. doi:10.1207/s15327876mp0401_1

McDaniel, M. A., Hartman, H. S., Whetzel, D. L., & Grubb, W. L. I. I. I. (2007). Situational judgement tests: Response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91. doi:10.1111/j.1744-6570.2007.00065.x

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640–647. doi:10.1037/0021-9010.75.6.640

Ones, D. S., & Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant datasets. *Journal of Applied Psychology, 83*, 35–42. doi:10.1037/0021-9010.83.1.35

Perlman, K., & Sanchez, J. I. (2010). Work analysis. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 73–98). New York, NY: Routledge.

Peterson, N., & Sager, C. E. (2010). The Dictionary of Occupational Titles and the Occupational Information Network. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 887–908). New York, NY: Routledge.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., . . . Dye, D. M. (2001). Understanding work using the occupational information network (O*NET): Implications for practice and research. *Personnel Psychology, 54*, 451–492. doi:10.1111/j.1744-6570.2001.tb00100.x

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and research.* Mahwah, NJ: Erlbaum.

Renaissance Partnership for Improving Teacher Quality. (2001). *Project overview*. Retrieved from http://www.uni.edu/itq/ProjectOverview/index.htm

Robertson, I. T., & Downs, S. (1989). Work-sample tests of trainability: A meta-analysis. *Journal of Applied Psychology, 74*, 402–410. doi:10.1037/0021-9010.74.3.402

Robertson, I. T., & Mindel, R. M. (1980). A study of trainability testing. *Journal of Occupational Psychology, 53*, 131–138. doi:10.1111/j.2044-8325.1980.tb00017.x

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*, 1009–1037. doi:10.1111/j.1744-6570.2005.00714.x

Roth, P. L., Bobko, P., McFarland, L. A., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. *Personnel Psychology, 61*, 637–661. doi:10.1111/j.1744-6570.2008.00125.x

Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*, 694–706. doi:10.1037/0021-9010.88.4.694

Rupp, D. E., & Searles, R. (2011). Using assessment centers to facilitate collaborated, quasi-standard, industry wide selection: Lessons learned from medical specialty placement in England and Wales. In N. Povah & G. C. Thornton III (Eds.), *Assessment centres and global talent management* (pp. 209–223). Farnham, Surrey, England: Gower.

Russell, C. J., & Dean, M. A. (1994, August). *The effect of history on meta-analytic results: An example from personnel selection research.* Presented at the annual meeting of the Academy of Management, Dallas, TX.

Rynes, S. L. (1993). When recruitment fails to attract: Individual expectations meet organizational realities in recruitment. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment* (pp. 27–40). Hillsdale, NJ: Erlbaum.

Rynes, S. L., & Miller, H. E. (1983). Recruiter and job influences on candidates for employment. *Journal of Applied Psychology, 68*, 147–154. doi:10.1037/0021-9010.68.1.147

Salgado, J. F., Viswesvaran, C., & Ones, D. C. (2002). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology* (pp. 165–199). Thousand Oaks, CA: Sage.

Schalock, M. D. (1998). Accountability, student learning, and the preparation and licensure of teachers: Oregon's teacher work sample methodology. *Journal of Personnel Evaluation in Education, 12*, 269–285. doi:10.1023/A:1008023412335

Schmidt, F. L., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43*, 627–670. doi:10.1146/annurev.ps.43.020192.003211

Schmidt, F. L., & Hunter, J. E. (1998). The validity of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research

findings. *Psychological Bulletin*, *124*, 262–274. doi:10.1037/0033-2909.124.2.262

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 115–139). New York, NY: Wiley.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422. doi:10.1111/j.1744-6570.1984.tb00519.x

Schmitt, N., & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86*, 451–458. doi:10.1037/0021-9010.86.3.451

Shaw, J. D., Gupta, N., Mitra, A., & Ledford, G. E. (2005). Success and survival of skill-based pay plans. *Journal of Management, 31*, 28–49. doi:10.1177/0149206304271376

Taylor, M. S. (1988). Effects of college internships on individual participants. *Journal of Applied Psychology, 73*, 393–401. doi:10.1037/0021-9010.73.3.393

Thornton, G. C., III, & Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review, 19*, 169–187. doi:10.1016/j.hrmr.2009.02.002

Thornton, G. C., III, & Mueller-Hanson, R. A. (2004). *Developing organizational simulations: A guide for practitioners and students*. Mahwah, NJ: Erlbaum.

Thornton, G. C., III, & Rupp, D. R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Erlbaum.

Thornton, G. C., III, Wilson, C. L., Johnson, R. M., & Rogers, D. A. (2009). Managing assessment center practices in the context of employment discrimination litigation. *Psychologist Manager, 12*, 175–186. doi:10.1080/10887150903103422

Truxillo, D. M., Donahue, L. M., & Kuang, D. (2004). Work samples, performance tests, and competency testing. In J. C. Thomas (Ed.), *Comprehensive handbook of psychological assessment: Vol. 4. Industrial and organizational assessment* (pp. 345–370). Hoboken, NJ: Wiley.

U.S. Department of Labor. (1991). *Dictionary of occupational titles* (4th ed.). Washington, DC: U.S. Government Printing Office.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574. doi:10.1037/0021-9010.81.5.557

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679–700. doi:10.1111/j.1744-6570.1999.tb00176.x

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376. doi:10.1037/h0026244

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*, 188–202. doi:10.1016/j.hrmr.2009.03.007

CHAPTER 30

SITUATIONAL JUDGMENT MEASURES

Robert E. Ployhart and Anna-Katherine Ward

Measures of situational judgment are increasingly used in work and educational settings. Most situational judgment measures present respondents with a realistic scenario or situation and then ask them how they would or should handle the situation. The situations usually involve dilemmas or competing goals so that respondents are forced to make difficult judgments and decisions. Situational judgment measures have a long history in industrial and organizational psychology, but their popularity has exploded in the past 20 years. They are also increasingly being applied in educational contexts. However, despite widespread use in practice, many fundamental questions remain about how to best measure and assess situational judgment.

The purpose of this chapter is to review the science and practice of situational judgment measures and identify areas in need of further research. First, the chapter briefly describes the underlying theory and nature of situational judgment measures. Second, the validity and reliability of situational judgment measures are reviewed. Third, how situational judgment measures are usually developed, scored, and administered is summarized. Finally, the chapter identifies areas in which future research is warranted. In particular, we describe the need to develop homogeneous and construct-valid measures, the need for more dynamic process models of judgment, the impact of multimedia assessment methods, and cross-cultural assessment.

Note that many comprehensive reviews of situational judgment measures are available (Chan & Schmitt, 2005; Lievens, Peeters, & Schollaert, 2008; McDaniel & Nguyen, 2001; Schmitt & Chan, 2006) as well as an entire edited volume on situational judgment measures in organizational settings (Weekley & Ployhart, 2006). This focused review summarizes and builds on these earlier reviews.

SITUATIONAL JUDGMENT MEASURES IN PRACTICE

The chapter begins with a description of what situational judgment measures are and what they are not. Next how these measures have been used in different areas of psychology is illustrated.

To describe assessments of situational judgment, it is first important to make a distinction between latent constructs and manifest measures (Arthur & Villado, 2008; Ployhart, Schneider, & Schmitt, 2006, Chapter 7). *Latent constructs* are the types of knowledge, skill, ability, or other characteristics (KSAOs) of theoretical interest. They are not directly observable or measurable, so they must be inferred through manifest indicators (i.e., measures). There are many potential ways to measure any particular KSAO. For example, the personality trait conscientiousness may be measured by a self-report paper survey, an online implicit measure that uses reaction times, a conditional reasoning approach, an interview, or the perceptions of other observers.

This chapter emphasizes the distinction between latent constructs and manifest measures because situational judgment measures are widely recognized to be multidimensional measurement methods

DOI: 10.1037/14047-030
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

(Chan & Schmitt, 2005). There is no unidimensional latent construct assessed by past or contemporary situational judgment measures. Rather, as explained later, situational judgment measures actually assess a variety of latent constructs simultaneously (hence, their description as a multidimensional method). Moreover, the same situational judgment measurement methodology can be used to measure different latent constructs in different contexts.

The underlying logic of a situational judgment measure is that effective performance in many real-world contexts requires sound judgment (Brooks & Highhouse, 2006). Hastie and Dawes (2001) referred to judgment as one's ability to "infer, estimate, and predict the character of events" (p. 48). Individual differences in judgment are found, with these individual differences being formed by differences in experience, education, cognitive ability, and personality (Hastie & Dawes, 2001; Payne, Bettman, & Johnson, 1993). Thus, situational judgment is one's ability to make sound predictions or conclusions about the outcomes of behaviors in a given context. As implied by the name, situational judgment measures are to a large degree contextually bound. This does not mean they are situationally specific; to the extent that the context shares the same situational elements, the same situational judgment measure may be appropriate. For example, the validity of a situational judgment measure of customer service in one setting is likely (within sampling variability) to be similar to its validity in another setting, to the extent that the two settings share similar features (e.g., coworker dynamics, reward structures).

Many different types of situational judgment measures exist, but they all share some common features. First, they present respondents with realistic situations. Here, the term *realistic* is used simply to note that the situations presented to respondents are representative of situations drawn from some context of interest. For example, the relevant situations for a situational judgment measure designed to predict retail sales performance will be drawn from the retail organization's actual customer service situations (how this is done is described in the section Developing and Administering Situational Judgment Measures later in this chapter). Second, they will ask respondents to identify what they should do, would do, or think is appropriate in that situation. In this manner, situational judgment is not an open-ended question but rather a form of a multiple-choice assessment (or at least uses a constrained-response option format). In this manner, measures of situational judgment are different than situational interviews that present respondents with hypothetical questions and then allow them to freely respond in their own words how they might handle those situations (Ployhart & MacKenzie, 2010). They are also different from work samples because work samples provide respondents with identical work tasks and then have them actually perform the behaviors needed to accomplish the work task. Finally, situational judgment measures can be distinct from pure knowledge tests in that the focus is not on knowledge but on judgment, that is, the application of knowledge and critical analysis to make predictions (see Brooks & Highhouse, 2006).

Thus, because situational judgment measures tend to be more realistic than a more context-generic assessment (e.g., cognitive ability, personality) but less realistic than work samples and related simulations, situational judgment measures are often considered a low-fidelity simulation (Motowidlo, Dunnette, & Carter, 1990). An example situational judgment item for the job of grocery cashier follows:

> It is a slow time of the day, and you are operating the only cash register in a grocery store. Unexpectedly, your cash register quits working, and a manager comes over to fix it. While it is getting repaired, the line at your register keeps getting longer, and customers are increasingly upset and impatient. Suppose your manager has to run to his office to get a key to access the machine. What would you say to the customers in this situation?
>
> (a) Say nothing; it is better to remain silent.
>
> (b) "The manager is having difficulty fixing the register, so you will need to wait a few more minutes."

(c) "I'm sorry for the delay, but we'll get you through as fast as possible."

(d) "These machines always break down; don't you hate technology?"

Notice that in this example, the respondent must pick from a set of options that are themselves derived from realistic situations. However, situational judgment items have many different formats. Some situational judgment measures use shorter situations, and others use longer situations. Instructions can also vary such that some ask what one should do, what one would do, or what the correct answer is. Occasionally, a situational judgment measure is developed so that there is one broad situation and several sets of more specific follow-up situations and questions. Most situational judgment measures have respondents evaluate between four and six response options, but some measures use only two options. Finally, some situational judgment measures use forced-choice formats, and others use Likert-type response formats in which respondents provide a rating for each option. Weekley, Ployhart, and Holtz (2006) provided a detailed review of variations in situational judgment measure formats (see also Lievens et al., 2008; Ployhart & MacKenzie, 2010). However, the sample item presented here is perhaps the most commonly used type of situational judgment format. Table 30.1 summarizes the key dimensions on which situational judgment measures frequently differ.

CONSTRUCT VALIDITY AND RELIABILITY OF SITUATIONAL JUDGMENT MEASURES

The construct validity and reliability evidence for situational judgment measures are summarized next. Obviously, the way these measures are structured (i.e., instructions, format, scoring) can influence validity and reliability, but the data on different types of situational judgment methods is not sufficient to make comparisons across types of structure. Therefore, this summary is necessarily broad and treats situational measures somewhat generically.

Construct Validity

Situational judgment measures tend not to have construct validity in the traditional sense because they are measurement methods (Chan & Schmitt, 2005; Schmitt & Chan, 2006). Instead, one needs to consider whether the content and structure of the measure is an appropriate assessment for the constructs it is designed to assess (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Arthur & Villado, 2008). Weekley et al. (2006) reported many different dimensions on which situational judgment methods may differ,

TABLE 30.1

Elements Distinguishing Different Situational Judgment Items

Dimension	Representative examples and variations
Situation complexity	Relatively short, simple situations to complex, detailed situations
Response format	Multiple choice, true–false, constructed response (open ended), oral, verbal, behavioral enactment
Response instructions	Would do, should do, most or least appropriate, best, worst, Likert-type scales
Reading level	Irrespective of complexity, items can be written at low or high reading levels
Test length	Short (roughly five to 10 items) to approximately 100 items; most between 20 and 40 items
Item independence	Nonindependent (e.g., branching, where response to an item influences the administration of subsequent items) to independent
Homogeneity	Some tests written to target a single construct, but most a multidimensional composite of constructs
Scoring	A single correct answer, points for multiple correct answers, different points depending on the appropriateness of responses, penalties (loss of points) for choosing inappropriate responses, continuous (Likert-type) scores on an item
Media or presentation format	Paper and pencil, video (real media or computer-generated avatars), audio only, Web or smartphone applications

implying that differences exist which types of constructs are most amenable to different types of measurement structure (see also Table 30.1). Research has primarily examined the consequences of response instructions on construct validity, finding that there is some distinction between "would do" and "should do" instructions (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel & Nguyen, 2001; Ployhart & Ehrhart, 2003). Would-do instructions tend to correlate more strongly with personality, and should-do instructions tend to correlate more strongly with cognitive ability. However, the differences for criterion-related validity are negligible (Lievens, Sackett, & Buyse, 2009; McDaniel et al., 2007). Although important, these findings are far from compelling evidence that situational judgment structure influences construct validity, and more systematic research needs to manipulate situational judgment structure to see its effects on construct validity. For example, one might expect to see a lower relationship with cognitive ability when constructed-response formats are used (relative to multiple-choice formats).

The nature of construct relationships found with situational judgment measures may (to a degree) also differ across contexts (McDaniel & Nguyen, 2001). For example, agreeableness should be more strongly related to situational judgment measures in customer service settings, and cognitive ability should be more strongly related to situational judgment measures in managerial settings.

These caveats aside, the limited data available on the construct validity of situational measures is based on meta-analyses linking situational judgment measures with a host of individual-difference KSAOs. Even though most situational judgment measures are intended to assess leadership skills, social skills, teamwork skills, knowledge, or personality (see Christian, Edwards, & Bradley, 2010), the measures actually have little homogeneity. Lievens et al. (2008) and Ployhart and MacKenzie (2010) reviewed the literature on situational judgment measure construct validity. Their reviews showed that situational judgment measures are moderately related to cognitive ability, personality (based on the five-factor model), and knowledge. McDaniel et al. (2007) provided meta-analytic evidence for many of these relationships. The correlations corrected for methodological artifacts were cognitive ability, .32; knowledge, .26; and the five-factor model, ranging from .13 to 27. These findings are consistent with the theory presented earlier suggesting that situational judgment is a multidimensional construct determined by more homogeneous individual differences in ability, knowledge, and personality.

A recent alternative conceptualization of situational judgment measures has been proposed by Motowidlo, Hooper, and Jackson (2006a, 2006b). They argued that situational judgment measures are actually assessments of procedural knowledge. However, this procedural knowledge may relate not just to task knowledge but also to knowledge about how trait expression may be appropriate in different situations. They termed this latter knowledge *implicit trait policies*. For example, those who have higher standing on agreeableness will be more likely to identify behaviors manifesting agreeableness as more appropriate in a given situation. Motowidlo and Beier (2010) have shown that it is possible to develop situational judgment measures to assess the implicit trait policies and, in doing so, provide stronger inferences of construct validity. These results are encouraging, but more research is necessary to understand the construct validity of the implicit trait policy approach.

Overall, meta-analytic evidence has suggested that situational judgment measures assess several individual-difference KSAO constructs simultaneously, and yet the magnitude of these relationships suggest they are not redundant with these KSAOs. This evidence is consistent with the theory underlying situational judgment measures; sound judgment requires a composite of cognitive ability, personality, and knowledge. However, as Ployhart and MacKenzie (2010) noted, this meta-analytic evidence fails to address two of the key questions in the situational judgment literature: Can a situational judgment measure be designed to assess specific aspects of judgment (i.e., more homogeneous constructs), and what types of structural design features may influence the nature of these construct relationships? These questions are addressed at the end of this chapter.

Reliability

Within the situational judgment measure literature, there is some ambiguity about the appropriate forms of reliability. Most have argued that internal consistency reliability is inappropriate given the multidimensional nature of situational judgment (i.e., the latent construct is not homogeneous). For example, McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) reported internal consistency estimates ranging from .43 to .94 (M = .60), but then later McDaniel et al. (2007) did not report internal consistency reliability because they claimed it was inappropriate to do so. To our knowledge, no successful attempts have been made to develop construct-homogeneous situational judgment measures; consequently, it would seem inappropriate to use estimates of internal consistency reliability. To illustrate, a technical report prepared by one of the authors of this chapter (Ployhart, 2008) summarized internal consistency reliabilities for nearly all situational judgment measures published up until 2008. He found these estimates ranged from .26 to .85, but the latter measures were based on composites and large numbers of items. In general, the number of items correlated .42 with internal consistency reliability, which is likely an underestimate because the relationship is not linear (additional information on internal consistency reliability may be found in Chapter 2, this volume).

Other research has examined the challenges of creating alternate forms in situational judgment measures. Lievens and Sackett (2007) found that different ways of creating alternative forms (e.g., randomly sampling vs. item cloning) result in differences in internal consistency reliability. Simply randomly sampling items will not create homogeneous composites. This is perhaps to be expected, given that situational judgment measures are multidimensional. Oswald, Friede, Schmitt, Kim, and Ramsay (2005) and Lievens and Sackett (2007) presented many issues that need to be addressed when developing item clones and alternate forms.

Many scholars have, however, argued that test–retest reliability is the more appropriate form of reliability for situational judgment measures (Lievens et al., 2008; McDaniel et al., 2007; Motowidlo et al., 1990). Schmitt and Chan (2006) described a few studies reporting test–retest reliability and suggested that these estimates are typically, but not always, higher. These estimates can range from the .20s to the .90s, but most cluster at .60 and higher. Perhaps more important, Schmitt and Chan (2006) posed the question of how large test–retest reliabilities should be, given that they are partly dependent on experience that is clearly malleable and changing over time. Of course, other ways of estimating reliability for multidimensional measures or composites are available, and these forms of reliability should be explored with situational judgment measures.

Applications of Situational Judgment Measures

Measures of situational judgment have become widely used in employment personnel selection settings since the early 1990s. They are now being applied in other domains such as education, certification testing, and training and development.

Personnel selection. The vast majority of research on situational judgment measures has been conducted within the context of personnel selection (see Lievens et al., 2008, for a comprehensive review). Early versions of situational judgment measures have existed since the late 1800s and have been used with varying degrees of popularity ever since (see DuBois, 1970; McDaniel et al., 2001). However, the birth of the contemporary situational judgment measurement was stimulated by Motowidlo et al. (1990). They presented the situational judgment measure as a low-fidelity simulation and showed that it had relatively strong validity (uncorrected correlations approximately .30), thus rivaling many of the most valid selection predictors available. Shortly thereafter, Chan and Schmitt (1997) showed how video-based situational judgment measures can have considerably smaller racial subgroup differences than written situational judgment measures, suggesting that higher fidelity measures of situational judgment may be a means for reducing adverse impact.

The Motowidlo et al. (1990) and Chan and Schmitt (1997) studies sparked considerable interest in situational judgment measures for selection purposes because they have high validity but smaller

subgroup differences. More recent meta-analyses have largely supported these earlier findings. McDaniel et al. (2007) found the correlation between measures of situational judgment and overall job performance to be .26 (corrected; the uncorrected correlation is .20). Christian et al. (2010) found similar validities but further showed that the validities can differ by the constructs intended to be measured (note the small number of studies for some of the constructs in this meta-analysis make it difficult to compare across constructs). Likewise, Ployhart and Holtz (2008) reported that Whites score approximately 0.5 standard deviation higher than Blacks, Hispanics, and Asians. These differences are smaller than those found for cognitive ability (with the exception that Asians tend to score higher than Whites on cognitive ability). Sex differences favor women but are small (approximately 0.10 standard deviation). Thus, measures of situational judgment help balance validity with diversity.

However, it must be emphasized that nearly all of this research has been conducted on job incumbents. Indeed, the meta-analysis by McDaniel et al. (2001) identified only six studies conducted with job applicants, and the criterion-related validity in those studies was only half as large as that found in incumbent settings. More recent research has further found that mean scores tend to be lower in applicant settings (MacKenzie, Ployhart, Weekley, & Ehlers, 2009). These results are quite different from the usual finding with noncognitive predictors (e.g., personality) showing that applicants score higher than incumbents. MacKenzie et al. (2009) also found that inferences of construct validity were partially affected by context, such that the relationships with cognitive ability were stronger in incumbent samples (no differences were found for relationships with personality).

Test takers often view situational judgment measures favorably because they present realistic work-related situations (as opposed to more generic assessments such as cognitive ability and personality; e.g., Chan & Schmitt, 1997). In our experience, they also tend to get more buy in by hiring managers, partly because they help develop the measure (see the section Developing and Administering Situational Judgment Measures later in this chapter) and partly because they are created using realistic situations. Indeed, the realism of the situational measure may also contribute to it's serving as a realistic preview of the job. Bauer and Truxillo (2006) provided a detailed discussion of how applicants may perceive situational judgment measures.

Thus, over the past 20 years, situational judgment measures have found widespread application in practice. Nearly every major vendor of personnel selection instruments offers a measure of situational judgment. Moreover, many larger vendors have developed situational judgment measures for different occupations (e.g., managerial, customer service). One potentially major limitation of this work is that nearly the entire literature base is composed of concurrent validities. Nevertheless, as a result of their success in the employment selection field, they are now being used in other areas.

Education. Interest is growing in exploring the usefulness of noncognitive predictors for making college admissions decisions. This interest stems in part because of the public's negativity toward standardized ability testing for admissions purposes. Situational judgment measures are a particularly appealing option, relative to other noncognitive predictors such as personality, because they have greater face validity and may also be more aligned to the criterion domain (cf. Lievens, Buyse, & Sackett, 2005). Work in this area is just emerging, but to date the results have been encouraging. Lievens et al. (2009) showed that situational judgment measures could predict academic GPA in a medical student population (uncorrected correlations approximately .15). In an undergraduate sample, Oswald, Schmitt, Kim, Ramsay, and Gillespie (2004) similarly found a criterion-related validity of .16 (uncorrected) with GPA and also a correlation of −.27 with absenteeism. More important, they also found that the situational judgment measure provided incremental validity over more cognitive-oriented standardized tests (SAT and ACT), yet had smaller subgroup differences. There appear to be opportunities to supplement traditional college admissions processes with situational measures (see also Cullen, Sackett, & Lievens, 2006; Schmitt et al., 2007). That said, a number of practical concerns are associated with

this suggestion, such as the effects of coaching and preparation on situational judgment scores and challenges (e.g., cost, difficulty) with item generation, cloning, and equivalence.

Certification testing. Certification testing differs from personnel selection testing in many ways, but perhaps the most important difference is in terms of their focus (American Educational Research Association et al., 1999). Certification testing emphasizes whether a candidate passes some threshold on the underlying latent construct. If a candidate meets or passes the threshold, she or he is deemed to be competent or capable, and no limit is set on the number of people who may meet or pass the threshold. In contrast, personnel selection emphasizes the rank ordering of applicants, such that those applicants who score highest will be hired first. Cut scores are frequently used to establish a threshold, but unlike certification testing, they are used primarily to screen out large numbers of candidates to make the process more manageable.

Because of the multidimensional nature of situational judgment measures, it can be challenging to use them for certification purposes. First, situational measures tend to have lower estimates of internal consistency reliability than standardized tests of knowledge, skill, or ability, making it difficult to set precise cut scores. Second, situational measures do not assess homogeneous constructs, making it difficult to isolate the specific threshold in a given domain. For example, it is possible that some people manifest sound judgment because of their experience, and others do so because of greater cognitive ability. There are multiple ways to reach the same judgments in a given context, so it can be difficult to set cut scores that are fair to all respondents. These challenges aside, it is possible to use situational judgment measures for certification. For example, the Canadian Council of Human Resources Associations uses a situational judgment measure for human resource certification. One of the potential benefits of this approach is that the assessment is seen as more realistic and hence more acceptable to test takers, who consist largely of working adults.

Training and development. Situational judgment measures are just starting to be used as tools for training and development. Although no published studies have used these measures in this manner, Fritzsche, Stagl, Salas, and Burke (2006) presented several technical reports and conference papers in which such situational judgment measures have been used effectively for training evaluation purposes. Moreover, they developed an agenda for integrating situational judgment measures into the broader literature and science on scenario-based training. The authors of this chapter have themselves developed situational judgment measures (a) to identify training needs among personnel, (b) as an evaluation of training effectiveness, and (c) as a diagnostic tool for development. For example, they have used situational judgment measures as a means to gauge one's understanding of ethical issues in human resources. They have also developed situational judgment measures to be used as tools for training in cross-cultural sensitivity. For example, trainees first learn about cultural differences between their host country and those of the target country. They are then tested to see whether they can assimilate to the dominant cultural manifestations of good judgment in work-related situations. To date, these results have been encouraging, but considerably more research is necessary.

DEVELOPING AND ADMINISTERING SITUATIONAL JUDGMENT MEASURES

The chapter now addresses how situational judgment measures are usually developed, scored, and administered. Because they are measurement methods, there is much variation in the way situational judgment measures are developed. However, little is known about how these structural features may influence validity and reliability. Weekley et al. (2006) summarized most of the available literature and presented an agenda for future research on design features. Most of their implications are summarized here.

The most common development procedure for situational judgment measures follows a three-step process. The first step involves sampling from the situational domain those situations that are relatively important and common but difficult and challenging to handle (Ployhart & MacKenzie, 2010).

That is, one adopts a domain-sampling approach to situations. Usually, one will use the critical incident method to record these situations (see Motowidlo et al., 1990; Weekley & Jones, 1999). The critical incident method is an approach in which participants identify the antecedent of a behavior, the behavior itself, and then the context within which it occurs. However, it is possible to identify job-relevant situations simply by meeting with experts and having them describe the context of their work. These experts are normally job incumbents who are currently performing the job, and they are asked to describe the situations as they actually occurred (not as how they might wish they had occurred). Regardless of how it is done, what is most important is that the situation domain is exhaustively sampled and is representative of the actual context. The following are samples of situations from customer service contexts that have to do with impression management (taken from Ployhart, 1999):

> A customer is telling you a story that is not particularly interesting to you. It is a long story, and you have several other customers who need help.
>
> You are helping a customer who keeps talking and talking. You have more customers to help.
>
> A customer tells a joke that you don't find the least bit funny. The customer seems surprised that you are not laughing.

The second step involves identifying appropriate and inappropriate behaviors for each situation. For example, in each of the situations just noted, job incumbents will describe how they or other employees might handle the situation. They are asked to record all relevant types of behavioral responses, both effective and ineffective, which will often produce a list of four to eight response options for each situation. However, these options are often redundant with each other, and so the final list is usually reduced to four. It is important to note that, again, one should ask respondents to report what is actually done, not how it should be done. One continues with this step until there is some redundancy and consistency in the behavior response options identified. At that point, the researcher will edit the final list to the desired number of response options (usually four or five per situation).

The final step involves creating the scoring key. This step differs from the first two steps in that supervisors, rather than job incumbents, are used. Supervisors may simply identify the most appropriate response option, identify the most and least appropriate response options, or rate each option on its effectiveness. For no apparent reason other than precedent (Motowidlo et al., 1990), most developers score the items such that respondents must pick the most and least appropriate behavioral response options for that situation. If the respondent correctly chooses the most appropriate option, she or he will receive 1 point. If the respondent correctly chooses the least appropriate option, she or he will again receive 1 point. If the respondent incorrectly picks the worst option as the best, she or he will lose 1 point. If the respondent incorrectly picks the best option as the worst, she or he will lose 1 point. If the respondent picks neither the best nor worst options, she or he will receive no points. Thus, item scores range from -2 to 2. However, this is just one approach, and depending on the goals and purpose of the measure, it is appropriate to use other scoring methods (e.g., Likert, single correct answer, modal response); see Weekley et al. (2006). However, no published research has suggested that one scoring system is preferable to another, and such research is sorely needed.

Earlier in this chapter, it was noted that there is a distinction in the instructions used for situational judgment measures (see McDaniel et al., 2007). Some measures use would-do instructions for each item, and others use should-do instructions. Research conducted in high-stakes field settings has found little difference between the two instruction formats in terms of criterion-related validity (although some modest differences have been found in construct validity; see Lievens et al., 2009; McDaniel et al., 2007).

SITUATIONAL JUDGMENT RESEARCH NEEDS

This review of situational judgment measures concludes by identifying areas in which future research

is particularly needed. These research needs stem from an understanding of the literature and from experience as practitioners.

Structure for Validity

Weekley et al. (2006) argued that research should examine how changing the development and scoring of situational judgment measures might influence construct validity, criterion-related validity, and reliability. Research has been done on should-do versus would-do instructions, but almost none has been done relating to other aspects of situational judgment measure structure. This is a bit curious, considering the abundance of research on structure relating to other predictor methods such as interviews and assessment centers. If one broadly summarizes those literatures, it is clear that structure determines the nature of the constructs assessed, which simultaneously means that the constructs one wishes to assess should determine structure. The only published attempt we found that adopts this approach is Motowidlo and Beier (2010), who attempted to develop situational judgment measures of implicit trait policies. Other approaches need consideration, and there are many interesting possibilities.

First, we echo the sentiments of Brooks and Highhouse (2006), who questioned why the situational judgment research has included no serious consideration of judgment. Indeed, it is rare in situational judgment research to even define the nature of the intended constructs to be assessed, much less show any construct validity evidence. Many researchers are guilty of this failing. However, if one truly believes that situational judgment is a construct representing individual differences in one's ability to predict events and reach conclusions about situations (Hastie & Dawes, 2001), then this conceptualization should drive the measure's structure, development, and scoring. The obvious place to start is for scholars to become familiar with the more basic research on judgment and decision making and to incorporate this theory into their design of situational judgment measures (cf. Brooks & Highhouse, 2006). Indeed, it would be refreshing to see more theoretical attention paid to situational judgment as a construct and process rather than as a sole focus on measurement. For example, scholars could adopt the methods and designs of judgment and decision-making research (e.g., cognitive tracing, cognitive task analysis, think-aloud protocols) to determine how respondents complete situational judgment measures. That said, perhaps situational judgment measures are not so much measures of judgment as they are measures of other individual differences such as cognitive ability and personality—and if so, then including *judgment* in the label of such measures is misleading.

Second, we continue to believe that research must identify whether it is possible to create situational judgment measures that can assess homogeneous constructs, or at least homogeneous subsets within a heterogeneous overall composite. This involves more than just obtaining greater understanding of construct validity (although that is certainly a worthy goal!); it also involves understanding how one can structure situational judgment measures to target specific constructs. It is dangerous practice to administer a test when one cannot convincingly explain what it measures and why it may be related to criteria of interest, yet test constructors do this all the time with measures of situational judgment. Understanding how to structure situational judgment measures to assess specific constructs would also enhance internal consistency reliability and possibly reduce hurdles and concerns about using these measures in high-stakes contexts such as licensure, credentialing, and educational testing.

However scholars choose to study situational judgment structure, we encourage them to think big and not manipulate very minute aspects of structure. Instead, we believe such research will only proceed in a productive manner if it tackles structural elements capable of producing big effects. Video versus paper-and-pencil comparisons represent one such major element, as does branching versus independent item formats. Researchers should draw from findings in the situational interview literature and the simulation literature to identify the factors that are likely to produce the largest effects on validity rather than simply vary all possible combinations of structure.

Process Models of Judgment

The basic judgment and decision-making literature has suggested that judgments do not occur in a vacuum but are influenced by context and may evolve dynamically over time (Hastie & Dawes, 2001; Payne et al., 1993). Sound judgment should change as information about contingencies, situational features, and experience evolves, which implies that judgment is not a static event but a process that occurs over time (even if time is measured in milliseconds). Brooks and Highhouse (2006) briefly introduced these issues and noted that situational judgment items may differ in their response latencies, depending on whether the judgments are more intuitive (fast) or analytical (slow). Yet, studies of the psychological response process underlying situational judgment are rare, even though such process models are recognized as a form of validity evidence in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999).

Drawing on psychological response process models from the survey and attitude literatures (Krosnick, 1999; Tourangeau, Rips, & Rasinski, 2000), Ployhart (2006) offered a process model for situational judgment measures. The predictor response process model posits that interpreting and responding to a situational judgment item requires four distinct processes: comprehension, retrieval, judgment, and response. Individual-difference KSAOs, such as cognitive ability and personality, have differential relationships to these different processes. For example, cognitive ability is likely related to how well (and quickly) one comprehends and retrieves from memory relevant examples of appropriate behavior. However, the judgment and response may also be driven by individual differences in personality. The predictor response process model represents just one approach to understanding the cognitive processes that underlie situational judgment measures, but it could be used to merge the literature on judgment and decision making into the situational judgment literature to assist in the design of experiments informing construct validity. For example, respondents may be asked to read the situation, and then some participants could also complete an interference task as they try to retrieve from memory relevant examples of effective behavior. Thus, by decomposing the psychological response process into cognitive operations, it becomes possible to apply the experimental methods of cognitive psychology to understand situational judgment construct validity.

Multimedia Methodologies

Since Motowidlo et al. (1990), attempts have been made to adapt the paper situational judgment methodology into a multimedia format. For example, Chan and Schmitt (1997) showed how administering a situational judgment measure using a video-based format can reduce racial subgroup differences (because of reduced reading requirements). Lievens and Sackett (2006) further showed that a video-based situational judgment measure can have higher criterion-related validity (for GPA and interpersonal criteria) than a written version. A further benefit of higher fidelity situational judgment measures is that they usually engender more favorable reactions (Chan & Schmitt, 1997; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000). However, prior research has barely scratched the surface of the technological opportunities for making situational judgment measures more realistic and higher fidelity. It is now possible to develop incredibly real simulations that can be administered and scored quickly and efficiently over the Internet. Situational judgment measures may be adapted such that they no longer require a multiple-choice type of format but instead allow respondents to behave by playing the game. Indeed, one of the authors of this chapter helped develop a Web-based situational judgment measure in which respondents had to adapt dynamically to changes in the situation and broader context (e.g., Lozzi, Cracraft, McKee, Ployhart, & Zaccaro, 2004). They were never administered multiple-choice questions but instead manifested their behaviors directly via their actions with the mouse. These technological opportunities create incredible operational and psychometric challenges because they can generate mountains of data that may or may not produce scores relevant to the construct or criterion.

Cross-Cultural Assessment

Most situational judgment measure research has been conducted in the United States, but these

tests are now being applied all around the globe. However, Lievens (2006) discussed a variety of issues that may arise when a situational judgment test is applied in a culture in which it was not designed. To begin, cultural differences may affect the transportability of item characteristics (see Volume 3, Chapter 26, this handbook). For example, the same situation and behavior may be linked to a particular construct in one culture but to a completely different construct in another culture. In other words, the item-construct relationship may vary from culture to culture. In addition, judgment is influenced by cultural norms and values. For example, MacKenzie, Ployhart, and Weekley (2007) found modest relationships between Schwartz's (1992) cultural values and situational judgment test responses. Therefore, what is judged to be the appropriate response choice by subject matter experts in the United States may not correspond with what is considered to be appropriate in another country. For instance, Nishii, Ployhart, Sacco, Wiechmann, and Rogg (2001) found that respondents from countries that value individualism deem item choices involving task orientation and direct communication to be the most appropriate, whereas those from more collectivistic cultures preferred options focusing on group harmony and saving face.

Lievens (2006) also discussed the problem of attempting to match predictor and criterion domains across cultures. He emphasized that using a situational judgment measure in a different culture than the one in which it was developed is comparable to applying it to a job for which it was not specifically designed. If the predictor (situational judgment test) is developed in the United States (as currently is typically the case), but criterion-related data are gathered in another country, the data do not necessarily represent the criterion for which the predictor is designed. This "imposed etic" approach (Berry, 1969) assumes that existing techniques can be applied regardless of culture, which may not be the case with situational judgment measures. For example, Such and Schmidt (2004) found that a situational judgment test based on a cross-cultural job analysis was valid in the United Kingdom and Australia (which are culturally similar countries), but not in Mexico (which is culturally different from the other two).

This is not to say that situational judgment tests cannot be used effectively in a variety of cultures; rather, one must be cautious as to where and how they are developed and validated. Lievens (2006) suggested that situational judgment measures that are cognitively loaded may show less variance across cultures, because cognitive constructs, in general, are more culturally invariant. Also, situational judgment measures can be developed and applied within the specific culture being targeted. For example, Bank and Latham (2009) found that a situational judgment test developed in Iran and based on Middle Eastern values and scenarios relevant to Iranian culture was a valid predictor of job performance in a population of Iranian employees. This emic approach (Berry, 1969), in which a test is developed and validated in a single country, has been successful in Asia (e.g., Chan & Schmitt, 2002) and in Europe (e.g., Born, 1994; Funke & Schuler, 1998; Lievens & Coetsier, 2002).

CONCLUSION

Situational judgment measures represent an increasingly popular methodology in selection contexts. They are also starting to be applied in educational testing, certification, and training and development. Although industrial and organizational psychology has learned much about situational judgment measures in the past 20 years, most of this work has been limited to attempts to determine whether they can have criterion-related validity and the types of constructs with which they are associated. In terms of directions for future research, we suggest researchers take a more active investigation of situational judgment by adopting experimental methods to understand their validity, using technology to improve the user experience and create administrative efficiencies, and understanding their applicability across cultures. The field has much to learn about situational judgment measures, but one might suspect that future research will also lead to many exciting developments in practice and application.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435–442. doi:10.1037/0021-9010.93.2.435

Bauer, T. N., & Truxillo, D. M. (2006). Applicant reaction to situational judgment tests: Research and related practical issues. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 233–249). Mahwah, NJ: Erlbaum.

Berry, J. (1969). On cross-cultural comparability. *International Journal of Psychology, 4*, 119–128. doi:10.1080/00207596908247261

Born, M. P. (1994). Development of a situation-response inventory for managerial selection. *International Journal of Selection and Assessment, 2*, 45–52. doi:10.1111/j.1468-2389.1994.tb00128.x

Brooks, M. E., & Highhouse, S. (2006). Can good judgment be measured? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 39–55). Mahwah, NJ: Erlbaum.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159. doi:10.1037/0021-9010.82.1.143

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233–254. doi:10.1207/S15327043HUP1503_01

Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, O. Smit-Voskuyl, & N. Anderson (Eds.), *The handbook of selection* (pp. 219–242). Oxford, England: Blackwell.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83–117. doi:10.1111/j.1744-6570.2009.01163.x

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14*, 142–155.

DuBois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn & Bacon.

Fritzsche, B. A., Stagl, K. C., Salas, E., & Burke, C. S. (2006). Enhancing the design, delivery, and evaluation of scenario-based training: Can situational judgment tests contribute? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 301–318). Mahwah, NJ: Erlbaum.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6*, 115–123. doi:10.1111/1468-2389.00080

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making.* Thousand Oaks, CA: Sage.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567. doi:10.1146/annurev.psych.50.1.537

Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 183–279). Mahwah, NJ: Erlbaum.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442–452. doi:10.1037/0021-9010.90.3.442

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment, 10*, 245–257. doi:10.1111/1468-2389.00215

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37,* 426–441. doi:10.1108/00483480810877598

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91,* 1181–1188. doi:10.1037/0021-9010.91.5.1181

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92*, 1043–1055. doi:10.1037/0021-9010.92.4.1043

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology, 94,* 1095–1101. doi:10.1037/a0014628

Lozzi, D. E., Cracraft, M., McKee, S., Ployhart, R. E., & Zaccaro, S. (2004, March). *Adaptive leadership: Assessing adaptive leadership through a new measurement technique.* Paper presented at the annual mid-year conference of the Engineering and Military Psychology Divisions, Ft. Belvoir, VA.

MacKenzie, W., Ployhart, R. E., & Weekley, J. A. (2007, April). *The relationship between culture and situational judgment responses*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, New York, NY.

MacKenzie, W. I., Ployhart, R. E., Weekley, J. A., & Ehlers, C. (2009). Contextual effects on SJT responses: An examination of construct validity and mean differences across applicant and incumbent contexts. *Human Performance, 23,* 1–21. doi:10.1080/08959280903400143

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91. doi:10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740. doi:10.1037/0021-9010.86.4.730

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103–113.

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*, 321–333. doi:10.1037/a0017975

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640–647. doi:10.1037/0021-9010.75.6.640

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*, 749–761.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57–81). Mahwah, NJ: Erlbaum.

Nishii, L. H., Ployhart, R. E., Sacco, J. M., Wiechmann, D., & Rogg, K. L. (2001). *The influence of culture on situational judgment test responses.* Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. H., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. *Organizational Research Methods, 8,* 149–164. doi:10.1177/1094428105275365

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187–207. doi:10.1037/0021-9010.89.2.187

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge, England: Cambridge University Press

Ployhart, R. E. (1999). *Integrating personality with situational judgment for the prediction of customer service performance.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 83–105). Mahwah, NJ: Erlbaum.

Ployhart, R. E. (2008). *Review of the National Professional Practice Assessment*. Report submitted to the Canadian Council of Human Resource Associations, Ottawa, Ontario, Canada.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11*, 1–16. doi:10.1111/1468-2389.00222

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172. doi:10.1111/j.1744-6570.2008.00109.x

Ployhart, R. E., & MacKenzie, W. (2010). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Vol. 2. Selecting and developing members for the organization* (pp. 237–252). Washington, DC: American Psychological Association.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and research.* Mahwah, NJ: Erlbaum.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85,* 880–887. doi:10.1037/0021-9010.85.6.880

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135–155). Mahwah, NJ: Erlbaum.

Schmitt, N., Oswald, F. L., Kim, B. H., Imus, A., Merritt, S., Friede, A., & Shivpuri, S. (2007). The use of background and ability profiles to predict college student outcomes. *Journal of Applied Psychology, 92,* 165–179. doi:10.1037/0021-9010.92.1.165

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). New York, NY: Academic Press.

Such, M. J., & Schmidt, D. B. (2004, April). *Examining the effectiveness of empirical keying: A cross-cultural perspective*. Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679–700. doi:10.1111/j.1744-6570.1999.tb00176.x

Weekley, J. A., & Ployhart, R. E. (Eds.). (2006). *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ: Erlbaum.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182). Mahwah, NJ: Erlbaum.

CHAPTER 31

HOLISTIC ASSESSMENT FOR SELECTION AND PLACEMENT

Scott Highhouse and John A. Kostek

Holism in assessment is a school of thought or belief system rather than a specific technique. It is based on the notion that assessment of future success requires taking into account the whole person. In its strongest form, individual test scores or measurement ratings are subordinate to expert diagnoses. Traditional standardized tests are seen as providing only limited snapshots of a person, and expert intuition is viewed as the only way to understand how attributes interact to create a complex whole. Expert intuition is used not only to gather information but also to properly execute data combination. Under the holism school, an expert combination of cues qualifies as a method or process of measurement. For example, according to Ruscio (2003), "Holistic judgments are premised on the notion that interactions among all of the information must be taken into account to properly contextualize data gathered in a realm where everything can influence everything else" (p. 1). The holistic assessor views the assessment of personality and ability as an ideographic enterprise, wherein the uniqueness of the individual is emphasized and nomothetic generalizations are downplayed (see Allport, 1962). This belief system has been widely adopted in college admissions and is implicitly held by employers who rely exclusively on traditional employment interviews to make hiring decisions. Milder forms of holistic belief systems are also held by a sizable minority of organizational psychologists—ones who conduct managerial, executive, or special-operation assessments.

In this chapter, the roots of holistic assessment for selection and placement decisions are reviewed and the applications of holistic assessment in college admissions and employee selection are discussed. Evidence and controversy surrounding holistic practices are examined, and the assumptions of the holistic school are evaluated. That the use of more-standardized procedures over less standardized ones invariably enhances the scientific integrity of the assessment process is a conclusion of the chapter.

HISTORICAL ROOTS

The traditional testing and measurement tradition is associated with people such as Sir Francis Galton and James McKeen Cattell (see DuBois, 1970, for a review). The holistic assessment tradition for selection and placement, however, was developed by psychologists outside of this circle. The intellectual forefathers of holistic assessment were influenced by gestalt concepts and were concerned with personality diagnosis for the purposes of selecting officers and specialists during World War II. The most prominent of these were Max Simoneit of Germany, W. R. Bion of England, and Henry A. Murray of the United States.

Max Simoneit

Max Simoneit was the chief of German military psychology during World War II. The Germans believed that victory depended on the superior leadership and intellect of the officer (Ansbacher, 1941). Simoneit, therefore, believed that psychological diagnosis (i.e., character analysis) of officer candidates and specialists should be the primary focus of

DOI: 10.1037/14047-031
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

military psychology. Assessments were qualitative rather than quantitative and subjective rather than objective (Burt, 1942). Simoneit believed that intelligence assessment was inseparable from personality assessment (Harrell & Churchill, 1941) and that an officer candidate needed to be observed in action to assess his total character. Although little is known about Simoneit, it is believed that he studied under the psychologist Narziss Ach (Ansbacher, 1941). Ach believed that willpower could be studied experimentally using a series of nonsense syllables as interference while a subject attempted to produce a rhyme (Ach, 1910/2006). As with Ach, assessment of will power was a central theme in Simoneit's work (Geuter, 1992). He devised tests such as obstacle courses that could not be completed and repeated climbs up inclines until the candidate was beyond exhaustion (Harrell & Churchill, 1941). These tests were accompanied by diagnoses of facial expressions, handwriting, and leadership role-plays. Simoneit's methods were seen as innovative, and the use of multiple and unorthodox assessment methods inspired officer selection practices used in Australia, Britain, and the United States (Highhouse, 2002).

W. R. Bion

W. R. Bion was trained as a psychoanalyst in England and became an early pioneer of group dynamics (Bion, 1959). He was enlisted to assist the war effort by developing a method to better assess officers and their likelihood of success in the field. According to Trist (2000), the British War Officer Selection Board was using a procedure in which psychiatrists interviewed officer candidates, and psychologists administered a battery of tests. This procedure created considerable tensions concerning how much weight to give to psychiatric versus psychological conclusions. Bion replaced this process with a series of leaderless group situations—inspired by the German selection procedures—to examine the interplay of individual personalities in a social situation. Bion believed that presenting candidates with a leaderless situation (e.g., a group carrying a heavy load over a series of obstacles) indicated their capacity for mature social relations (Sutherland & Fitzpatrick, 1945). More specifically, Bion believed that the pressure for the candidate to look good individually was put into competition with the pressure for the candidate to cooperate to get the job done. The challenge for the candidate was to demonstrate his abilities through the medium of others (Murray, 1990). Candidates underwent a series of tests and exercises over a period of 2.5 days. Psychiatrists and psychologists worked together as an observer team to share observations and develop a consensus impression of each candidate's total personality.

Henry A. Murray

Henry A. Murray was originally trained as a physician but quickly abandoned that career when he became interested in the ideas of psychologist Carl Jung. He developed his own ideas about holistic personality assessment while working as assistant director, and later director, of the Harvard Psychological Clinic in the 1930s. During the war, Murray was enlisted by the Office of Strategic Services (OSS) to develop a program to assess and select future secret agents. Murray's medical training involved grand rounds, in which a team of varied specialists contributed their points of views in arriving at a diagnosis. He believed that one shortcoming of clinical case studies was that they were produced by a single author rather than a group of assessors working together (Anderson, 1992). Accordingly, Murray and his colleagues assembled an OSS assessment staff that included clinical psychologists, animal psychologists, social psychologists, sociologists, and cultural anthropologists. Conspicuously absent from his team were personnel psychologists (Capshew, 1999). Murray developed what he called an *organismic* approach to assessment. The approach, described in detail in *Assessment of Men* (OSS, 1948), involved multiple assessors inferring general traits and their interrelations from a number of specific signs exhibited by a candidate engaged in role plays, simulations, group discussions, and in-depth interviews—and combining these inferences into a diagnosis of personality. Murray's procedures were the inspiration for modern-day assessment centers used for selecting and developing managerial talent (Bray, 1964).

The three figures discussed in this section were mavericks who rejected the prevailing wisdom that consistency is the key to good measurement.

Examiners were often encouraged to vary testing procedures from candidate to candidate and to give special attention to tests they preferred. In other words, there was little appreciation for the concepts of reliability and standardization. Although many celebrated the fresh approach brought about by the holistic pioneers, others questioned the appropriateness of many of their practices (Eysenck, 1953; Older, 1948).

APPLICATIONS IN SELECTION AND PLACEMENT

Much has been written on the application of holistic principles in clinical settings (see Grove, Zald, Lebow, Snitz, & Nelson, 2000; Korchin & Schuldberg, 1981), but their application to selection and placement decisions has received considerably less attention (cf. Dawes, 1971; Ganzach, Kluger, & Klayman, 2000; Highhouse, 2002). It is notable, however, that one of the earliest debates about the use of holistic versus analytical practices involved the employee selection decision-making domain (Freyd, 1926; Viteles, 1925). Morris Viteles (1925) objected to the then-common practice of making decisions about applicants on the basis of test scores alone. According to Viteles,

> It must be recognized that the competency of the applicant for a great many jobs in industry, perhaps even for a majority of them, cannot be observed from an objective score any more than the ability of a child to profit from one or another kind of educational treatment can be observed from such a score. (p. 134)

Viteles (1925) believed that the psychologist in industry must integrate test scores with clinical observations. According to Viteles, "His judgment is a diagnosis, as that of a physician, based upon a consideration of all the data affecting success or failure on the job" (p. 137). Max Freyd (1926) responded that psychologists are unable to agree, even among themselves, on a person's abilities by simply observing the person. Rather than diagnosing a job candidate, Freyd argued that the psychologist should make subjective impressions objective by incorporating them into a rating scale. According to Freyd,

> The psychologist cannot point to the factors other than test scores upon which he based his correct judgments unless he keeps a record of his objective judgments on the factors and compares these records with the vocational success of the men judged. Thus he is forced to adopt the statistical viewpoint. (p. 353).

Most modern-day organizational psychologists share Freyd's (1926) view of assessment for selection, but those who practice assessment at the managerial and executive level are less likely to do so (Jeanneret & Silzer, 1998; Prien, Schippman, & Prien, 2003).

The most common applications of holism in assessment and selection practice are discussed next. These include (a) college admissions decision making, (b) assessment centers, and (c) individual assessment.

College Admissions

Colleges and universities have continuously struggled with how to select students who will be successful while at the same time ensuring opportunity for underrepresented populations (see Volume 3, Chapters 14 and 15, this handbook, for more information on this type of testing). Standardized tests provide valuable information on a person's degree of ability to benefit from higher education. Admissions officers, however, are charged with ensuring a culturally rich and diverse campus and accepting students who will exhibit exceptional personal qualities such as leadership and motivation. In 2003, the U.S. Supreme Court (*Gratz v. Bollinger*, 2003) ruled that it is lawful for admissions decisions to be influenced by diversity goals, but that holistic, individualized selection procedures, not mechanical methods, must be used to achieve these goals (see McGaghie & Kreiter, 2005). This decision was in response to the University of Michigan's then practice of awarding points to undergraduate applicants based on, among other things, their minority status. These points were aggregated into an overall score, according to a fixed, transparent formula. Justice Rehnquist argued

that consideration of applicants must be done at the individual level rather than at the group level. Race, according to the majority decision, is to be considered as one of many factors, using a holistic, case-by-case analysis of each applicant. In his dissenting opinion, Justice Souter argued that such an approach only encourages admissions committees to hide the influence of (still illegal) racial quotas on their decisions (*Gratz v. Bollinger*, 2003).

In 2008, Wake Forest University became the first top-30 U.S. university to drop the standardized test requirement for undergraduate admissions. Wake Forest moved to a system in which every applicant is eligible for an admission interview (Allman, 2009). The Wake Forest interviews do not follow any specific format, and interviewers are free to ask different questions of different applicants. Although the Wake Forest interviewers make overall interview ratings on a scale ranging from 1 to 7, the admissions committee explicitly avoids using a numerical weight in the overall applicant evaluation (Hoover & Supiano, 2010). Wake Forest is a clear exemplar of what Cabrera and Burkum (2001) referred to as the holistic era of college admissions in the United States.

Assessment Centers

The notion that psychologists could select people for higher level jobs was not widely accepted until after World War II (Stagner, 1957). The practices used to select officers in the German, British, and U.S. militaries were seen as having considerable potential for application in postwar industry (Brody & Powell, 1947; Fraser, 1947; Taft, 1948). Perhaps most notable was Douglas Bray's assessment center (Bray, 1964). Bray, inspired by the 1948 OSS report *Assessment of Men*, put together a team of psychologists to implement a program of tests, interviews, and situational performance tasks for the assessment of the traits and skills of prospective AT&T managers. Although the original assessment center was used exclusively for research, the procedure evolved into operational assessment centers still in use today. Unlike the original, clinically focused center, the operational assessment centers of today focus on performance in situational exercises, and they commonly use managers as assessors. The focus on standardization and objective rating is in contrast to the earlier holistic practices advocated by the World War II psychologists (Highhouse, 2002). One similarity that remains between the modern and early assessment centers, however, is the use of rater consensus judgments. The consensus judgment process is predicated on the notion that observations of behavior must be intuitively integrated into an overall rating (Thornton & Byham, 1982). This consensus judgment process involves discussion of everyone's ratings to arrive at final dimension ratings and ultimately an overall assessment rating for each candidate. The group discussion process can take several days to complete and does not involve the use of mechanical or statistical formulas.

Individual Assessment

One area of managerial selection practice that has maintained the holistic school's emphasis on considering the whole person and intuitively integrating assessment information into a diagnosis of potential is commonly referred to as *individual assessment* (see Ryan & Sackett, 1992). Although the label is not very descriptive, it does emphasize the focus on idiographic (as opposed to nomothetic) assessment. Practices vary widely from assessor to assessor, but individual assessment typically involves intuitively combining impressions derived from scores on standardized and unstandardized psychological tests, information collected from unstructured and structured interviews, a candidate's work and family history, informal observation of mannerisms and behavior, fit with the hiring organization's culture, and fit with the job requirements. The implicit belief behind the practice is that the complicated characteristics of a high-level job candidate must be assessed by a similarly complicated human being (Highhouse, 2008). According to Prien, Shippmann, and Prien (2003), the holistic process of integration and interpretation is a "hallmark of the individual assessment practice" (p. 123; see also Ryan & Sackett, 1992).

Next, the evidence and controversy surrounding the use of holistic methods for making predictions about success in educational and occupational domains are reviewed. A summary of studies that have directly contrasted holistic versus analytical approaches in college admissions and employee selection is also provided.

EVIDENCE AND CONTROVERSY

The first study to empirically test the notion that experts could better integrate information holistically than analytically was conducted by T. R. Sarbin in 1942. Sarbin followed 162 freshmen who entered the University of Minnesota in 1939. Using a measure of 1st-year academic success, Sarbin compared the earlier prediction of admissions counselors with a statistical formula that combined high school rank and college aptitude test score. The counselors had access to these two pieces of information as well as information from additional ability, personality, and interest inventories. The counselors also interviewed the students before the fall quarter of classes. Of interest to Sarbin was the performance of the simple formula against the counselors' predictions. The results showed that the counselors, who had access to all of the test data and interview observations, did significantly worse in predicting 1st-year success than the simple (high school rank plus aptitude test score) formula. Subsequent studies on college admissions have supported the idea that a simple combination of scores is not only effective but is in many instances more effective than holistic assessment for predicting success in school.

Table 31.1 provides a summary of research comparing holistic to analytical approaches to college admissions. The table shows that in almost every case, holistic evaluations based on test scores, grades, and other personal evaluations (e.g., interviews, letters, biographical information) were equaled or exceeded by simple combinations of standardized tests scores and grades.

Organizational psychologists took note of findings like Sarbin's (1942), which were documented

TABLE 31.1

Empirical Comparisons of Holistic and Analytical Approaches to College Admissions

Source	Method	Results
Alexakos (1966)	Guidance counselors made predictions of college GPA on the basis of information collected from testing and interviews over a 4-year period.	The holistic judgments of counselors were slightly outperformed by a statistical combination of high school GPA, standardized test scores, and demographic variables.
Dawes (1971)	Psychology faculty predicted the success of incoming graduate students on the basis of a standardized test, undergraduate GPA, and letters of recommendation.	A mechanical model of the committee's judgment process predicted faculty ratings of graduate success better than the committee itself.
Hess (1977)	A medical school admissions committee predicted success in 1st-year chemistry on the basis of standardized tests, interviews, transcripts, and biographical data.	The holistic judgments of the committee did not predict success in 1st-year chemistry, whereas high school performance data alone were successful.
Rosen & Van Horn (1961)	A scholarship award committee predicted 1st-year GPA using high school rank, standardized test scores, biographical information, and letters of recommendation.	The holistic judgment of the award committee was equal to the use of only high school rank in predicting 1st-semester GPA in college.
Sarbin (1943)	College admissions officers predicted success on the basis of high school rank, a standardized test, and an intensive interview.	The holistic judgments of the admissions officers were inferior to a simple combination of high school rank and standardized test score.
Schofield (1970)	A medical school admissions committee predicted success on the basis of college GPA, a standardized test, biographical data, and letters of reference.	The holistic judgment of the admissions committee was equal to a statistical combination of only college GPA and standardized test scores.
Watley & Vance (1964)	Guidance counselors made predictions of college GPA and participation in activities, using high school rank, standardized tests, and biographical information.	The holistic judgments of counselors equaled a mechanical formula that included high school rank and test scores.

Note. Only studies using actual counselors, faculty, or admissions officers as assessors are included in this table. GPA = grade point average.

in Paul Meehl's classic 1954 book *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Moreover, early organizational studies seemed to support Meehl's findings that holistic integration of information was not living up to the claims of the personality assessment pioneers (e.g., Huse, 1962; Meyer, 1956; Miner, 1970). However, an influential review of judgmental predictions in executive assessment—dismissing the relevance of this controversy to the organizational arena (Korman, 1968, p. 312)—eased the mind of many industrial psychologists involved in assessment practice. Also, assessment center research was showing impressive criterion-related validity, suggesting that an approach with many subjective components could be quite useful in identifying effective managers (Howard, 1974).

Table 31.2 provides a summary of research comparing holistic to analytical approaches to selection and placement in the workplace. Although the studies vary in rigor and sometimes do not provide fair comparisons of holistic and analytical approaches, some broad inferences can be drawn from this compilation:

- There are surprisingly few studies on the relative effectiveness of holistic assessment for employee selection, especially as it regards individual assessment.
- Only one study clearly favored holistic assessment (i.e., an assessor with knowledge of a cognitive ability test score did better than the score alone; Albrecht, Glaser, & Marks, 1964), compared with at least five that clearly favored analytical approaches and at least seven that were a draw.
- The few studies to examine the incremental validity of holistic judgment have not provided encouraging results.

Our summary shows that evidence for the superiority of holistic judgment is quite rare in educational and employment settings. A meta-analysis comparing clinical to statistical predictions in primarily medical and health diagnosis settings found that statistical methods were at least equal to clinical methods in 94% of the cases and significantly superior to them in as much as 47% of studies (Grove et al., 2000). Despite the fact that clinicians often had access to more information than the formulas, the statistical methods were estimated to be approximately 10% better in overall accuracy.

Recall that advocates of the holistic school have suggested that experts may take into account the interactions among various pieces of assessment evidence and understand the idiosyncratic meaning of one piece of information within the context of the entire set of information for one candidate (Hollenbeck, 2009; Jeanneret & Silzer, 1998; Prien et al., 2003). Given such expertise, one might expect that holistic judgments—which consider all of the information at hand—should unequivocally outperform dry formulas based on ratings and test scores. This has not been the case.

The existing research on selection and placement decision making has provided disappointingly little evidence that subjectivity and intuition provide added value. Traditional employment interviews provide negligible incremental validity over standardized tests of cognitive ability and conscientiousness (Cortina, Goldstein, Payne, Davison, & Gilliland, 2000; see Chapter 27, this volume, for more information on employee interviews). Research has also unequivocally shown that the more the interview is structured or standardized to look like a test, the greater its utility for predicting on-the-job performance (Conway, Jako, & Goodman, 1995; McDaniel, Whetzell, Schmidt, & Maurer, 1994). Having assessors spend several days discussing job candidates in assessment centers, and arriving at an overall consensus rating for each, provides no advantage over taking a simple average of each person's ratings (Pynes, Bernardin, Benton, & McEvoy, 1988). Assessing a candidate's fit with the job—a common practice in individual assessment—also appears to provide little advantage in predicting a candidate's future job performance (Arthur, Bell, Villado, & Doverspike, 2006). Taken together, this research has suggested that considering each candidate as a unique prediction situation has not resulted in demonstrably better prediction. As Grove and Meehl (1996) noted in their review of the debate between ideographic versus nomothetic views of prediction: "That [debate] is clearly an empirical question rather than a purely philosophical one decidable from the armchair, and empirical

TABLE 31.2

Empirical Comparisons of Holistic and Analytical Approaches to Employee Selection and Placement

Source	Method	Results
Albrecht, Glaser, and Marks (1964)	Psychologists ranked managers on the basis of an intensive interview, cognitive ability tests, and projective tests.	The holistic judgments of psychologists outperformed the cognitive ability test score alone.
Borman (1982)	Military recruiters provided assessment center exercise effectiveness ratings and consensus overall assessment ratings.	A mechanical combination of unit-weighted exercise ratings slightly outperformed the holistic discussion-based judgments.
Feltham (1988)	Assessors provided exercise scores and consensus overall assessment ratings in a police assessment center.	A unit-weighted composite of exercise scores outperformed the holistic discussion-based judgments.
Ganzach, Kluger, and Klayman (2000)	Judgments of interviewers from the Israeli military were used as predictors of military transgressions.	Adding a holistic interviewer rating to mechanically combined interview dimension ratings slightly increased the prediction of the criterion.
Huse (1962)	Psychologists made final ratings on the basis of an intensive interview and standardized and projective tests.	The validities of holistic ratings based on complete data were not higher than validities based solely on standardized (paper-and-pencil) tests.
Meyer (1956)	Manager judgments were made on the basis of interview and standardized test scores.	Four of the five validity coefficients for holistic judgments were below the validity of a cognitive ability test alone.
Mitchel (1975)	Assessors provided overall potential ratings on the basis of exercise performance and test scores.	The multiple correlation of the predictors strongly outperformed the holistically derived overall assessment, but the two converged over cross-validation.
Pynes, Bernardin, Benton, and McEvoy (1988)	Assessors provided preconsensus and postconsensus dimension ratings and overall consensus ratings in a police assessment center.	The mechanically and holistically derived dimension ratings were indistinguishable ($r = .83$) and correlated strongly with the overall holistic judgment ($r = .71$ for both).
Roose & Doherty (1976)	Manager judgments were made on the basis of 64 cues from personnel files, including test, biographical, and objective interview data.	The mean increase in R^2 achieved by adding the holistic combination of cues by the judges over a linear combination of cues was 0.7%.
Sackett & Wilson (1982)	Assessors provided preconsensus and postconsensus dimension ratings and overall consensus ratings in a managerial assessment center.	A simple average of dimension ratings predicted postdiscussion ratings 93.5% of the time.
Trankell (1959)	One psychologist made predictions of Swedish airline pilot success in training on the basis of observations and standardized test scores.	The holistic evaluations slightly outperformed each of the test scores alone.
Tziner & Dolan (1982)	Assessors subjectively combined ratings, cognitive ability tests, and exercise ratings into an overall assessment.	The R of the predictors outperformed the holistically derived overall assessment.
Wollowick and McNamara (1969)	Assessors subjectively combined tests, dimension ratings, and exercise ratings into an overall assessment.	The R of the predictors strongly outperformed the holistically derived overall assessment.

evidence is, as described above, massive, varied, and consistent" (p. 310).

As such, the holistic approach to selection and placement as commonly practiced in hiring and admissions is not consistent with principles of evidence-based practice (Highhouse, 2002, 2008).

ASSUMPTIONS OF HOLISTIC ASSESSMENT

Given that the early promise of the holistic approach has not held up to scientific scrutiny, it is reasonable to ask why many people continue to hold this point

of view. Some common assumptions held by holistic assessors are outlined here.

Assessors Can Take Into Account Constellations of Traits and Abilities

Advocates of holistic assessment have argued that the expert combination of information is a sort of nonlinear geometry that is not amenable to standardization in some sort of simple formula (Prien et al., 2003, p. 123). This argument implies that holistic assessment is a sort of mystical process that cannot be made transparent. Ruscio (2003) compared it with the arguments of astrologers who, when faced with mounds of negative scientific evidence, reverted to whole-chart interpretations to render their professional judgments. Aside from the logical inconsistencies involved with the claim that assessors can take into account far more unique configurations of data than can be cognitively processed by humans, considerable evidence has shown that simple linear models perform quite well in almost all prediction situations faced by assessors (e.g., Dawes, 1979).

It has long been recognized that it is possible to include trait configurations in statistical formulas (e.g., Wickert & McFarland, 1967). However, very little research on the effectiveness of doing so in selection settings has been conducted, likely because predictive interactions are quite rare. Dawes (1979) noted that relations between psychological variables and outcomes tend to be monotonic. In contrast to conventional wisdom, nonmonotonic interactions (e.g., certain types of leaders are really good in one situation and really bad in another situation) are quite rare. Furthermore, the evidence has suggested that assessors could not make effective use of such interactions, even if they existed.

Assessors Can Identify Idiosyncrasies That Formulas Ignore

Meehl (1954) described the "broken leg case" in which a rare event may invalidate a prediction made by a formula. Meehl used the example of predicting whether Professor X would go to the cinema on a particular Friday night. A formula might take into account whether the professor goes to the movie on rainy or sunny days, prefers romantic comedies to action movies, and so forth. The formula may not, however, take into account the fact that Professor X broke his leg on the previous Monday. A human assessor could take into account such broken-leg cues. Although the example is compelling and is commonly used to justify the use of holistic assessment procedures, evidence has not supported the usefulness of broken-leg cues (see Camerer & Johnson, 1991). The problem seems to be that assessors overrely on idiosyncratic cues, not distinguishing the useful ones from the irrelevant ones. Assessors find too many broken legs.

Assessors Can Fine-Tune Predictions Made by Formulas

A related argument is the idea that assessors may use their experience and wisdom to modify predictions that are made mechanically (Silzer & Jeanneret, 1998). The problem with this argument is that it assumes that a prediction can be fine-tuned. As noted by Grove and Meehl (1996),

> If an equation predicts that Jones will do well in dental school, and the dean's committee, looking at the same set of facts, predicts that Jones will do poorly, it would be absurd to say, "The methods don't compete, we use both of them." (p. 300)

If a mechanical procedure determines that an executive is not suitable for a position as vice president, then fine-tuning the procedure involves overruling the mechanical prediction. Certainly, intuition could be used to alter the formula-based rank ordering of candidates. We have yet to find evidence that this results in an improvement in prediction of job performance.

Some Assessors Are Better Than Others

There are experts in many domains, but evidence for expertise in intuitive prediction is lacking. The renowned industrial psychologist Walter Dill Scott concluded long ago, "As a matter of fact, the skilled employment man probably is no better judge of men than the average foreman or department head" (Scott & Clothier, 1923, p. 26). Subsequent research on assessment centers has found few differences among assessors in validity (Borman, Eaton, Bryan, & Rosse, 1983). Similar findings have emerged for

the employment interview (Pulakos, Schmitt, Whitney, & Smith, 1996). After reviewing research on predictions made by clinicians, social workers, parole boards, judges, auditors, and admission committees, Camerer and Johnson (1991) concluded, "Training has some effects on accuracy, but experience has almost none" (p. 347). The burden of proof is on the assessor to demonstrate that he or she can predict better than someone with rudimentary training on the qualities important to the assessment.

Candidates for High-Level Jobs Do Not Differ Much on Ability and Personality

One common assumption of holistic assessment is that variability of test scores is restricted for people being selected at the highest levels of organizations. As one example, Stagner (1957) contended about executive assessment that

> simple, straightforward tests of intelligence and other objective measures seem not to have too much value, largely because an individual is not considered for such a position until he has already demonstrated a high level of aptitude in lower level activities. (p. 241)

Large-scale testing programs at Exxon and Sears in the 1950s, however, demonstrated that using a psychometric approach to identifying executive talent can be quite effective (Bentz, 1967; Sparks, 1990). Personality tests better predict behavior for jobs that provide more discretion (Barrick & Mount, 1993), and the validity of cognitive ability measures increases as the complexity of the job increases (Hunter, 1980). Research has also shown that managers and executives are more variable in cognitive ability than conventional wisdom would suggest (Ones & Dilchert, 2009). Test scores can predict for higher level jobs.

Formulas Become Obsolete

A final assumption to consider is the idea that formulas are static and inflexible and thus are not useful for making predictions about performance in the chaotic environments of the marketplace. According to Prien et al. (2003), "Economic conditions and circumstances and the nature of client businesses might be evolving, dynamic and in flux, changing so that any particular algorithm, no matter how carefully developed, could be obsolete" (p. 128). The problem with this argument is that assessors are somehow assumed to be more flexible and attuned to subtle changes in effectiveness criteria. In fact, assessors are likely to rely on implicit theories developed from past training and experience. Moreover, these implicit theories have likely become resistant to change as a result of positive illusions and hindsight biases (Fisher, 2008). Formulas may be updated on the basis of new information and empirical research.

FINAL THOUGHTS

As noted by Hogarth (1987), people's intuitive judgments are based on information processed and transformed by their minds. Hogarth noted that there are four ways in which judgments may derail (see Table 31.3): (a) selective perception of information, (b) imperfect information processing, (c) limited capacity, and (d) biased reconstruction of events.

Although humans have limited resources to make judgments, they paradoxically cope with this by adding more complexity to the problem. For example, people often create elaborate stories to make sense of disparate pieces of information, even when the stories themselves are too elaborate to be predictive (Gilovich, Griffin, & Kahneman, 2002; Pennington & Hastie, 1988). This chapter has shown that assessors are not immune to the limitations of human judgment. Indeed, assessment experience may only serve to exacerbate issues such as professional biases and overconfidence (Sieck & Arkes, 2005).

One benefit of the holistic school of thought is that it encourages people to look more broadly at the predictors and the criteria: to consider what the person brings to the educational or work environment as a whole. This broader perspective may encourage one to more thoroughly examine noncognitive attributes of the candidate, along with nontask attributes of job performance. The research does not, however, support the use of a holistic approach to data integration. Assessors are still needed to select the data on which the formulas are based and to assign ratings to the data points that are subjective in nature

TABLE 31.3

Four Ways That Assessor Judgments May Derail

Derailer	Explanation
Selective perception	Assessors are bombarded with information and must selectively choose which things deserve attention. People often see what they expect to see on the basis of initial impressions, background information, or professional biases.
Imperfect processing	Assessors cannot simultaneously integrate even the information on which they choose to focus. The sequence in which a person processes information may bias the judgment.
Limited capacity	Assessors use simple heuristics or rules of thumb to reduce mental effort. Comparing a candidate with oneself, or with people who have already succeeded in similar jobs, are strategies commonly used to simplify assessment.
Biased reconstruction	Memory is formed by assembling fragments of information. Vivid information is more easily recalled, even when it is not representative of the set as a whole. Also, information that supports one's initial impression is more easily recalled than information contrary to it.

(e.g., interpersonal warmth in the interview). If assessors heed the advice of Freyd (1926) by making subjective impressions objective, assessment will move out of the realm of philosophy, technique, and artistry and into the realm of science.

References

Ach, N. (2006). *On volition* (T. Herz, Trans.). Leipzig, Germany: Verlag von Quelle & Meyer. (Original work published 1910) Retrieved from http://www.psychologie.uni-konstanz.de/forschung/kognitive-psychologie/various/narziss-ach

Albrecht, P. A., Glaser, E. M., & Marks, J. (1964). Validation of a multiple-assessment procedure for managerial personnel. *Journal of Applied Psychology, 48*, 351–360.

Alexakos, C. E. (1966). Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions. *Journal of Educational Psychology, 57*, 297–306.

Allman, M. (August, 2009). *Quintessential questions: Wake Forest's admission director gives insight into the interview process.* Retrieved from http://rethinkingadmissions.blogs.wfu.edu/2009/08

Allport, G. W. (1962). The general and the unique in psychological science. *Journal of Personality, 30*, 405–422. doi:10.1111/j.1467-6494.1962.tb02313.x

Anderson, J. W. (1992). The life of Henry A. Murray: 1893–1988. In R. A. Zucker, A. I. Rabin, J. Aronoff, & S. J. Frank (Eds.), *Personality structure in the life course: Essays on personology in the Murray tradition* (pp. 304–334). New York, NY: Springer.

Ansbacher, H. L. (1941). German military psychology. *Psychological Bulletin, 38*, 370–392. doi:10.1037/h0056263

Arthur, W., Bell, S. T., Villado, A. J., & Doverspike, D. (2006). The use of person–organization fit in employment decision making: An assessment of its criterion-related validity. *Journal of Applied Psychology, 91*, 786–801. doi:10.1037/0021-9010.91.4.786

Barrick, M. R., & Mount, M. K. (1993). Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. *Journal of Applied Psychology, 78*, 111–118. doi:10.1037/0021-9010.78.1.111

Bentz, V. J. (1967). The Sears experience in the investigation, description, and prediction of executive behavior. In F. R. Wickert & D. E. McFarland (Eds.), *Measuring executive effectiveness* (pp. 147–205). New York, NY: Appleton-Century-Crofts.

Bion, W. R. (1959). *Experiences in groups: And other papers.* London, England: Tavistock.

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67*, 3–9.

Borman, W. C., Eaton, N. K., Bryan, J. D., & Rosse, R. L. (1983). Validity of Army recruiter behavioral assessment: Does the assessor make a difference? *Journal of Applied Psychology, 68*, 415–419. doi:10.1037/0021-9010.68.3.415

Bray, D. W. (1964). The management progress study. *American Psychologist, 19*, 419–420. doi:10.1037/h0039579

Brody, W., & Powell, N. J. (1947). A new approach to oral testing. *Educational and Psychological Measurement, 7*, 289–298. doi:10.1177/001316444700700206

Burt, C. (1942). Psychology in war: The military work of American and German psychologists. *Occupational Psychology, 16*, 95–110.

Cabrera, A. F., & Burkum, K. R. (2001, November). *Criterios en la admisión de estudiantes universitarios en los EEUU: Balance del sistema de acceso a la universidad (selectividad y modelos alternativos)* [College admission criteria in the United States: Assessment of the system of access to the university (selectivity and alternative models)]. Madrid, Spain: Universidad Politécnica de Madrid.

Camerer, C. F., & Johnson, E. J. (1991). The process–performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). Cambridge, England: Cambridge University Press.

Capshew, J. H. (1999). *Psychologists on the march: Science, practice, and professional identity in America, 1929–1969*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511572944

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565–579. doi:10.1037/0021-9010.80.5.565

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*, 325–351. doi:10.1111/j.1744-6570.2000.tb00204.x

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26*, 180–188. doi:10.1037/h0030868

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582. doi:10.1037/0003-066X.34.7.571

DuBois, P. H. (1970). *A history of psychological testing*. Boston, MA: Allyn & Bacon.

Eysenck, H. J. (1953). *Uses and abuses of psychology*. Harmondsworth, England: Penguin Books.

Feltham, R. (1988). Assessment centre decision making: judgemental vs. mechanical. *Journal of Occupational Psychology, 61*, 237–241.

Fisher, C. D. (2008). Why don't they learn? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 364–366. doi:10.1111/j.1754-9434.2008.00065.x

Frazer, J. M. (1947). New-type selection boards in industry. *Occupational Psychology, 21*, 170–178.

Freyd, M. (1925). The statistical viewpoint in vocational selection. *Journal of Applied Psychology, 9*, 349–356.

Ganzach, Y., Kluger, A. N., & Klayman, N. (2000). Making decisions from an interview: Expert measurement and mechanical combination. *Personnel Psychology, 53*, 1–20. doi:10.1111/j.1744-6570.2000.tb00191.x

Geuter, U. (1992). *The professionalization of psychology in Nazi Germany* (R. J. Holmes, Trans.). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511666872

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, NY: Cambridge University Press.

Gratz v. Bollinger, 539 U.S. 244 (2003).

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323. doi:10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction. *Psychological Assessment, 12*, 19–30. doi:10.1037/1040-3590.12.1.19

Harrell, T. W., & Churchill, R. D. (1941). The classification of military personnel. *Psychological Bulletin, 38*, 331–353. doi:10.1037/h0057675

Hess, T. G. (1977). Actuarial prediction of performance in a six-year AB-MD program. *Journal of Medical Education, 52*, 68–69.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55*, 363–396. doi:10.1111/j.1744-6570.2002.tb00114.x

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 333–342. doi:10.1111/j.1754-9434.2008.00058.x

Hogarth, R. M. (1987). *Judgment and choice: The psychology of decision*. New York, NY: Wiley.

Hollenbeck, G. P. (2009). Executive selection—What's right . . . and what's wrong. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 130–143. doi:10.1111/j.1754-9434.2009.01122.x

Hoover, E., & Supiano, B. (2010, May). *Admissions interviews: Still an art and a science*. Retrieved from http://chronicle.com/article/The-Enduring-Mystery-of-the/65545

Howard, A. (1974). An assessment of assessment centers. *Academy of Management Journal, 17*, 115–134. doi:10.2307/254776

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity*

generalization to the General Aptitude Test Battery (GATB). Washington, DC: U.S. Employment Service.

Huse, E. G. (1962). Assessments of higher-level personnel: The validity of assessment techniques based on systematically varied information. *Personnel Psychology, 15*, 195–205. doi:10.1111/j.1744-6570.1962.tb01861.x

Jeanneret, R., & Silzer, R. (1998). An overview of psychological assessment. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational settings* (pp. 3–26). San Francisco, CA: Jossey-Bass.

Korchin, S. J., & Schuldberg, D. (1981). The future of clinical assessment. *American Psychologist, 36*, 1147–1158. doi:10.1037/0003-066X.36.10.1147

Korman, A. K. (1968). The prediction of managerial performance: A review. *Personnel Psychology, 21*, 295–322. doi:10.1111/j.1744-6570.1968.tb02032.x

McDaniel, M. A., Whetzell, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616. doi:10.1037/0021-9010.79.4.599

McGaghie, W. C., & Kreiter, C. D. (2005). Special article: Holistic versus actuarial student selection. *Teaching and Learning in Medicine, 17*, 89–91. doi:10.1207/s15328015tlm1701_16

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press. doi:10.1037/11281-000

Meyer, H. H. (1956). An evaluation of a supervisory selection program. *Personnel Psychology, 9*, 499–513. doi:10.1111/j.1744-6570.1956.tb01082.x

Miner, J. B. (1970). Executive and personnel interviews as predictors of consulting success. *Personnel Psychology, 23*, 521–538. doi:10.1111/j.1744-6570.1970.tb01370.x

Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60*, 573–579.

Murray, H. (1990). The transformation of selection procedures: The War Office Selection Boards. In E. Trist & H. Murray (Eds.), *The social engagement of social science: A Tavistock anthology* (pp. 45–67). Philadelphia: University of Pennsylvania Press.

Office of Strategic Services. (1948). *Assessment of men*. New York, NY: Rinehart.

Older, H. J. (1948). [Review of the book *Assessment of men: Selection of personnel for the Office of Strategic Services*]. *Personnel Psychology, 1*, 386–391.

Ones, D. S., & Dilchert, S. (2009). How special are executives? How special should executive selection be? Observations and recommendations. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 163–170. doi:10.1111/j.1754-9434.2009.01127.x

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 521–533. doi:10.1037/0278-7393.14.3.521

Prien, E. P., Schippmann, J. S., & Prien, K. O. (2003). *Individual assessment: As practiced in industry and consulting*. Mahwah, NJ: Erlbaum.

Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychology, 49*, 85–102. doi:10.1111/j.1744-6570.1996.tb01792.x

Pynes, J., Bernardin, H. J., Benton, A. L., & McEvoy, G. M. (1988). Should assessment center dimension ratings be mechanically-derived? *Journal of Business and Psychology, 2*, 217–227. doi:10.1007/BF01014039

Roose, J. E., & Doherty, M. E. (1976). Judgment theory applied to the selection of life insurance salesmen. *Organizational Behavior and Human Decision Processes, 16*, 231–249.

Rosen, N. A., & Van Horn, J. W. (1961). Selection of college scholarship students: Statistical vs clinical methods. *Personnel Guidance Journal, 40*, 150–154.

Ruscio, J. (2003). Holistic judgment in clinical practice: Utility or futility? *Scientific Review of Mental Health Practice, 2*, 38–48.

Ryan, A. M., & Sackett, P. R. (1992). Relationships between graduate training, professional affiliation, and individual psychological assessment practices for personnel decisions. *Personnel Psychology, 45*, 363–387. doi:10.1111/j.1744-6570.1992.tb00854.x

Sackett, P. R., & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. *Journal of Applied Psychology, 67*, 10–17.

Sarbin, T. L. (1942). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48*, 593–602. doi:10.1086/219248

Schofield, W. (1970). A modified actuarial method in the selection of medical students. *Journal of Medical Education, 45*, 740–744.

Scott, W. D., & Clothier, R. C. (1923). *Personnel management.* Chicago, IL: Shaw.

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making, 18*, 29–53. doi:10.1002/bdm.486

Silzer, R., & Jeanneret, R. (1998). Anticipating the future: Assessment strategies for tomorrow. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological*

assessment: Predicting behavior in organizational settings (pp. 445–477). San Francisco, CA: Jossey-Bass.

Sparks, C. P. (1990). Testing for management potential. In K. E. Clark & M. B. Clark (Eds.), *Measures of leadership* (pp. 103–112). West Orange, NJ: Leadership Library of America.

Stagner, R. (1957). Some problems in contemporary industrial psychology. *Bulletin of the Menninger Clinic, 21*, 238–247.

Sutherland, J. D., & Fitzpatrick, G. A. (1945). Some approaches to group problems in the British Army. *Sociometry*, *8*, 205–217. doi:10.2307/2785043

Taft, R. (1948). Use of the "group situation observation" method in the selection of trainee executives. *Journal of Applied Psychology, 32*, 587–594. doi:10.1037/h0061967

Thornton, G., & Byham, W. (1982). *Assessment centers and managerial performance.* New York, NY: Academic Press.

Trankell, A. (1959). The psychologist as an instrument of prediction. *Journal of Applied Psychology, 43*, 170–175.

Trist, E. (2000). Working with Bion in the 1940s: The group decade. In M. Pines (Ed.), *Bion and group psychotherapy* (pp. 1–46). London, England: Jessica Kingsley.

Tziner, A., & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67*, 728–736.

Viteles, M. S. (1925). The clinical viewpoint in vocational selection. *Journal of Applied Psychology, 9*, 131–138. doi:10.1037/h0071305

Watley, D. J., & Vance, F. L. (1964). *Clinical versus actuarial prediction of college achievement and leadership activity*. Minneapolis: University of Minnesota Student Counseling Bureau.

Wickert, F. R., & McFarland, D. E. (1967). *Measuring executive effectiveness.* New York, NY: Appleton-Century-Crofts.

Wollowick, H. B., & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology, 53*, 348–352.

CHAPTER 32

EMPLOYMENT TESTING AND ASSESSMENT IN MULTINATIONAL ORGANIZATIONS

Eugene Burke, Carly Vaughan, and Ray Glennon

Recognition of the value of scientifically valid assessment of the talents and potential of people applying for work at or already employed by private and public sector organizations has grown significantly in recent years. This increased recognition is partly the result of increased confidence in the evidence base supporting the criterion-related evidence of validity for cognitive ability tests and self-report questionnaires (e.g., Bartram, 2005; Hurtz & Donovan, 2000; Judge & Ilies, 2002; Ones, Dilchert, Viswesvaran, & Judge, 2007; Robertson & Smith, 2001; Salgado et al., 2003; Schmidt & Hunter, 1998). The importance of identifying and developing talent has also become more imperative across organizations and has acted to promote the use of tests and assessments in employment settings. In a recent review and synthesis of the literature on high potential, Silzer and Church (2009) stated that

> ever since the "war for talent" was popularized by the 1997 McKinsey study . . . the idea of identifying and managing high-potential talent has become increasingly important for organizations. At the very center of talent management . . . the singular ability to define and identify that elusive variable known as potential in an individual or group of individuals is considered a competitive advantage in the market place . . . today there is significant pressure on organizations and their leadership teams to ensure that they have well-validated and useful measures of potential. (pp. 377–378)

With increased globalization and labor mobility and the growth of the Internet, the context for employment testing and assessment has also changed significantly. Whereas testing programs might previously have been developed within a single national, cultural, and language context, many organizations now operate across national borders and require assessment solutions that meet their talent acquisition and talent development needs wherever they operate. Thus, the context for employment testing and assessment has become more complex. As House (2004) commented,

> The increasing connection among countries, and the globalization of corporations, does not mean that cultural differences are disappearing or diminishing. On the contrary, as economic borders come down, cultural barriers could go up, thus presenting new challenges and opportunities in business. (p. 5)

These connections have important measurement implications both within a business assessment context and beyond.

Consider a global leader in the retail sector that, having established itself as a market leader in Western Europe, is now expanding into Eastern Europe, Asia, and the United States. This organization is one that recognizes the value of people to its business and that talent acquisition (i.e., recruitment and selection) has a critical influence on the customer's experience in its stores and, therefore, on footfall, revenues, and sustaining the strength of its brand

DOI: 10.1037/14047-032
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

locally, nationally, and internationally. Consider next the choices facing this organization as it expands in its various geographic markets, including those that it might at one point have referred to as its home geography. On one hand, it recognizes that the diversity of local markets will need to be reflected in its local operations, including the people staffing its stores, yet on the other it also recognizes the need to maintain consistent standards for the people that it hires to manage and operate those stores, wherever they may be. So, how should this organization approach such issues, and as developers of assessment solutions and consultants on talent issues, how should testing professionals advise such a client?

This example typifies the type of assignment and assessment needs that today's globalized world now present to the assessment designer and consultant and to the manager concerned with acquiring and developing the talent required for the organization to be successful. The need to address diversity in terms of language, nationality, and culture is also true of mature markets in which employment testing and assessment are well established. In the late 1990s and early 2000s, for example, the United Kingdom saw substantial immigration from Eastern Europe by those seeking employment opportunities outside their country of origin. During that time, it was not unusual for us to receive requests from U.K. organizations operating solely within the United Kingdom for assessments in the languages of Eastern European countries in addition to British English. Thus, even within national boundaries, issues of whether a test or assessment is effective across multiple languages and cultures are likely to arise because of the simple fact that labor is now more mobile internationally.

As other countries and geographic regions emerge as economic powers, and as their economic development creates demand for talent management, and specifically talent assessment methodologies and processes, then similar challenges to those already alluded to will emerge within those geographic areas. So, the same questions will arise. For example, to what extent would these areas need to develop assessments that are specific to their languages and cultures, and, if they were to develop them, how would they gauge the effectiveness of those assessments compared with those used elsewhere in the world, particularly because economic competition is global, which in effect means that talent management is also global. Indeed, even within these emerging economies, issues of diversity in language and culture already exist—South Africa being an example of this point—and so the issues of addressing the challenge of linguistic and cultural issues in assessment can be seen as truly global and local or, to use a term that has become popular in the management literature, *glocal* in nature.

SCOPE AND FOCUS OF THIS CHAPTER

The original scope for this chapter was to focus on employment testing in multinational organizations. As just indicated, however, such a focus masks the fact that labor is as geographically mobile as are organizations, and the issues and solutions explored also apply to the need for tests and assessments that function within as well as across geographic areas. As such, the scope of this chapter was broadened to employment testing in multinational settings. The goal of this chapter is to share experiences and perspectives that will help inform the development of employment tests and assessments that are generalizable across populations, languages, and cultures and to offer knowledge from client engagement as well as product development in terms of processes and methods that can be used to guide testing and assessment from the original conception of a test or assessment through to evaluations of its effectiveness in a multinational context.

Before moving on, consider what is meant by *employment testing.* The most obvious manifestation of an employment test is in the form of an instrument assessing general and specific cognitive abilities that tends to be made up of items that have only one correct answer. These tests are also frequently referred to as *maximal performance measures,* given that one's score reflects how many correct responses were given across all the items in the test. In the more colloquial vernacular of industrial and organizational psychology, these tests are sometimes also referred to as assessments of people's "can-do" qualities.

Another frequently used form of employment test is the self-report questionnaire used to assess personality, motivation, and personal values. This class of test also includes integrity tests used to screen for counterproductive behaviors such as substance abuse and theft. Personality tests are also referred to as measures of typical performance because they are constructed to obtain information on the style of behavior that a person is likely to manifest, and therefore they explore the strength of a particular style rather than present items for which there is a single correct answer. Although maximal performance tests are seen as being related to assessments of people's can-do qualities, measures of personality, motivation, and values are sometimes referred to as measures of people's "will-do" qualities.

This chapter also covers a form of assessment that has become increasingly popular in both research and practice in employment testing, namely situational judgment tests (SJTs). Given that the title of this class of test includes the words *situation* and *judgment,* it, perhaps more than any other form of employment test, encapsulates the challenges of multinational testing and assessment. Such tests are widely used to assess how a person will respond to situations that are likely to occur in a job or role and can be used either as maximal performance measures when the task is to identify the most or least effective responses to a situation or as typical performance measures when the task is to select the answers that reflect the most or least typical response a person would give in that situation.

These are all examples of instruments or tests that are widely used in high-stakes assessments for preemployment screening, recruitment, selection, placement, and promotion of people within organizations. Their technical qualities therefore have real consequences for the people to whom these tests are administered and for the people who base employment decisions on the information such tests provide. Although this chapter does not focus on other forms of testing and assessment specifically, the experience and learning shared in this chapter are also relevant to them. Examples include simulations in the form of assessment center exercises (e.g., role plays and in-basket exercises) as well as instruments such as 360-degree appraisals (in which a person is rated on a number of dimensions by his or her superior, peers, direct reports, and others) that are widely used in lower stakes contexts such as personal development. Whether the test is used in a high- or low-stakes setting, the principal message is that its technical qualities must be shown to be adequate across the languages and cultures in which it is deployed. Given the wider application of the methods and processes that are described, we hope that the reader will not mind too much if we tend, on occasion, to use the terms *test* and *assessment* almost interchangeably.

Several core and fundamental issues in multinational testing and assessment are illustrated through case examples, that is, by taking a case and exploring it in terms of an issue that it helps to highlight. The first such case involves a client engagement with a global retailer and the development of an SJT to be deployed initially in Turkey, the United Kingdom, and the United States. For the reasons explained earlier, this class of test provides an opportunity to evaluate whether constructs are generalizable across language and cultural settings. No published research has examined whether an SJT can be deployed in a multinational context, and this challenge was particularly demanding in that one of the fundamental questions regarding SJTs is what they measure.

The second issue is that of *localization,* which is defined as a process that is more than just translation, although translation is obviously a key component. The reader may be surprised that this issue is not addressed first, but the case of the multinational deployment of an SJT serves to emphasize the need to be clear about the constructs that tests and assessments are designed to measure and to build into the process of test and assessment development those frameworks and procedures that evaluate upfront whether the constructs are generalizable and which linguistic and cultural factors need to be addressed early in constructing a solution to an assessment need.

Many texts on localization tend to focus on the statistical procedures and the evidence provided by those procedures on the psychometric equivalence of items, scales, and instruments across different languages and cultures. Acquiring the data required

by such procedures is an expensive and time-consuming undertaking, and should such data show technical inadequacies in any given language or culture, repairing the item, scale, or instrument adds to the cost of and delay in delivering an assessment solution. An examination of current practice can be used to develop a proactive series of process steps that mitigate against the risk of failure at the point of statistical checks. No claim is made that the process presented is the only solution to localization and adaptation of tests and assessments, but the evidence gathered to date has shown that it offers promise and that it does deliver efficiencies and cost savings.

The third issue addressed is that of norms. Norms are beginning to receive more attention but are still subject to confusion, particularly in the context of the multinational use of employment tests and assessments. Returning to the example of the global retailer, among the choices facing such a user of tests and assessments is to apply local national or language norms, international norms, or a combination of the two. In advising such a client, what data and relative benefits should a test provider or an assessment–talent consultant present and discuss with the client? This question is one example of many that need to be considered in addressing the issue of what norms are inappropriate in a multinational setting.

The fourth issue discussed is that of online delivery of assessments and test security. As much as the past decade or so has seen increased economic globalization, it has also seen the rapid growth and impact of the Internet on how businesses and organizations operate and how people in their personal lives carry out everyday transactions. In a world in which software-as-a-service models are widely used to drive business processes, and in which many of those applications are operated through cloud computing, the push for organizations to move their assessments online has increased markedly. The pull toward online testing and assessment is also strong when one considers the ease with which people now expect to execute transactions in their personal lives, such as purchasing, banking and investments, and even transactions with public bodies responsible for taxation and health. Thus, for reasons of speed, cost, the access to talent pools that it offers organizations, and the convenience it offers those seeking employment opportunities, Internet testing and assessment are growing. This growth presents a real challenge to the science and practice of valid assessment in how to safeguard the value of testing and assessment when such measures are administered under conditions in which a supervisor or a proctor is not physically present. It also makes test security a truly international issue in that Internet-delivered content can be accessed anywhere in the world.

Finally, trends that may become features of a future world of testing and assessment are explored (after all, every chapter such as this must have its crystal-ball moment). That section begins by referencing Samuel Messick's (1989, 1995, 1998) seminal work on validity as a framework for a general approach to developing and deploying employment tests and assessments in a multinational context (see also Chapter 4, this volume).

RECONSIDERING MESSICK AND THE ISSUES TO BE ADDRESSED IN PROVIDING EVIDENCE OF VALIDITY

Messick's (1989, 1995, 1998) work has continued to guide the measurement community's understanding of test validity. He originally proposed six facets of validity, which are briefly summarized here:

1. *content validity,* or evidence of content relevance, representativeness, and technical quality;
2. *substantive validity,* which refers to the theoretical rationale for respondents' observed responses and empirical evidence that the respondents engage in these processes;
3. *structural validity,* or evidence of the fidelity of the scoring structure to the construct domain;
4. *generalizability,* or the extent to which scoring properties and interpretations generalize across populations, settings, and tasks;
5. *external validity,* which incorporates convergent and discriminant evidence from multitrait and multimethod comparisons and criterion relevance and utility and which most closely relates to concerns regarding construct validity; and

6. *consequential validity,* which appraises value implications of score interpretations as a basis for action as well as the actual and potential consequences of test use, including bias, fairness, and distributive justice.

To these original six facets, Messick (1989) added a seventh in referring to *construct-irrelevant variance* in his later writing, a facet that gains in importance when one considers how a test or assessment can morph when it undergoes translation and localization into other languages. Indeed, when localizing self-report questionnaires into several languages, wording and context in the original language can present a significant barrier to maintaining construct and measurement equivalence. Test and assessment designers naturally tend to reflect the language and cultural biases of their backgrounds, and even minor biases in the form of a verb, adverb, noun (e.g., how many Americans own a conservatory?), or phrase (consider the difficulty of localizing the English phrase *tends to cut corners*) will present difficulties.

Why are these facets of validity important in the context of multinational testing? After all, is it not just a matter of translation? Consider the example of a multinational governmental agency that has deployed tests for the selection of staff across several national populations and has found that score differences are such that pass rates vary by nationality. The natural tendency may be to question whether the test or assessment is operating consistently across the different populations. However, it could also be that the people are indeed different and that this difference is reflected in the score differences observed. It could also be that variation occurred in the processes with which the test or assessment was administered (e.g., a recruitment advertisement that attracted applicants to the test or assessment session) or in the way in which administration was executed in different locations. These issues, and Messick's facets of validity, can be summed up in the form of three questions:

1. *Is it the test?* That is, does evidence exist that differences in scores across populations are the result of how the test was constructed and how it is scored? This question can be answered in part by looking at whether test items are biased against different populations, whether scoring keys operate in different ways across populations, whether scales fail to operate in an equivalent way across populations, or whether content such as the language and the context in which an item is seated and phrased has an impact on the generalizability of a construct as measured by a test or questionnaire.
2. *Is it the people?* It is entirely possible that, when evidence exists that a test or assessment is functionally and psychometrically equivalent across cultures, true differences across populations and national settings will be found. The question that naturally follows and should then be investigated is, what is it about the people that may be related to the differences observed in the scores? Are there demographic factors that are known to influence test scores? Finding such relationships may not, in itself, resolve the issues arising from test score differences, but identifying such differences may help to explain not only why scores differ but also whether a one-size-fits-all deployment of the tests will meet an organization's needs or whether that deployment would benefit from greater emphasis on the organization's local needs once the talent pools in different geographic areas are better understood. Either way, the differences may well have to do with the people rather than with the test.
3. *Is it the process?* Our experience in addressing the question of whether it is the people has often unveiled factors related to differences in practice and process across regional and national settings. For example, an organization's brand and its employer value proposition may vary in their appeal to talent in different geographic areas, something that often comes as a surprise to an organization used to its strong employer brand being recognized in those areas in which its presence is more established. As such, data may reveal a need for that organization to review its talent attraction and recruitment processes in those geographic areas in which it is not attracting the same bench strength of talent as in others. The possibility also exists that those responsible for, say, recruitment in certain

> geographic areas are operating by different criteria than those in others. In one case, we helped a client to understand that recruiters in some geographic areas were using criteria that considered both volume and quality of hire, whereas those in other areas were focusing on volume rather than the quality of hire attracted. So, data may reflect the simple fact that variations in processes within which the test or assessment is deployed can have a significant impact on the scores observed across different geographic areas and that such variations may have little to do with the quality of the test or assessment itself.

Although in practice variations in scores may be influenced by a combination of all three factors (tests, people, and processes), translating Messick's facets into these three more general considerations significantly benefits multinational employment testing from initial design through the evaluation of data gathered from the deployment of a test or assessment. Tests, people, and processes provide the key themes of this chapter, and we begin by considering the question "Is it the test?"—that is, whether the constructs that underpin a test or assessment to be deployed in a multinational setting are generalizable.

BUILDING GENERALIZABILITY INTO TESTING AND ASSESSMENT THROUGH THE COMBINED EMIC–ETIC APPROACH

House and Javidan (2004) defined a construct as referring to "the construction of conceptions or ideas by the investigator. A construct is the product of an investigator's creativity" (p. 20). Earlier in their overview of the Global Leadership and Organizational Behavior Effectiveness (GLOBE) research program, they pointed to a fundamental debate that has persisted in cross-cultural research and that has spread to the issue of multinational and cross-cultural employment testing and assessment:

> Two central aspects of cultures are frequently discussed in cross-cultural literature: culturally generalizable and culturally specific aspects. . . . A phenomenon is culturally generalizable if all cultures can be assessed in terms of a common metric and cultures can be compared in terms of such phenomena. In contrast, culturally specific phenomena occur only in a subset of cultures, and are not comparable across all cultures. (p. 19)

The GLOBE program exemplifies the issues and challenges of multinational employment testing and assessment. Is it possible to develop measures that are robust and operate across languages and cultures so that nations can be compared on various aspects of leadership? Here, then, is the basis of the emic–etic debate in cross-cultural research that stems from the views articulated by Pike in the 1950s (Berry, 1980; Geisinger, 2003; Pike, 1967) and from which the emic perspective, derived from *phonemic* and *phonemes,* argues that constructs are culturally bound and best understood from the perspective of the insider in the context of a given culture. In contrast, the etic perspective, derived from *phonetic,* argues that constructs can be culturally generalizable and objectively evaluated from the perspective of an outsider to a culture. As argued by Jahoda (1995), both perspectives play a role in developing tests and assessments that generalize to a multinational setting but, because this debate is inherently bound up with language, linguistic differences are often highlighted at the expense of the value of an etic perspective to the end user of the information provided by tests and assessments. Consider again the global retailer operating in several geographic areas, languages, and cultures. The extreme emic view would argue that each assessment should be developed from within the unique linguistic and cultural context of each geographic area. That approach could, in addition to the significant economic investment required, deliver to the organization a test or assessment that is unique to each context. The question that naturally arises is whether the test or assessment is operating with a common metric and equivalent constructs; although such issues are surmountable, such a solution adds considerable complexity in constructing and using tests and assessments developed in this way.

An example of what might be seen as an etic approach to this problem, and one that has been actual practice for many years, would be to export tests and assessments from a host language and culture, say an instrument developed in the United Kingdom or United States, where many employment tests originate. The assumption in such an approach is that the articulation of the constructs measured by the test or assessment is appropriate to the target language and culture. This may not be the case. The content can, of course, be adapted to attempt to overcome any shortcomings identified by various statistical analyses, but then the risk again arises of several iterations of adaption and the cost and time delay associated with executing those iterations.

Given the weaknesses of an either–or approach, the reader will not be surprised that a combined emic—etic approach is adopted here. This position is predicated on the following:

- That it is possible to develop and apply an etic taxonomy of work behaviors that generalizes across national, language, and cultural settings. Support for this approach comes from the work of those such as the GLOBE program already mentioned as well as the work of Hofstede (2001; Hofstede & Hofstede, 2005) and Schwartz (Schwartz, 1994; Schwartz & Bilsky, 1990; Schwartz & Boehnke, 2004) that has shown that it is possible to develop etic dimensions and models for constructs related to values and culture. Furthermore, as Hofstede's (2001; Hofstede & Hofstede, 2005) work has shown, such models are able to capture diversity within as well as between countries and are sensitive to organizational differences as well.
- That this taxonomy can be used to apply a criterion-centric approach to defining the behaviors that are critical to effective performance in a job or role irrespective of the geographic, language, or cultural setting in which that job or role needs to be performed. Bartram, Robertson and Callinan (2002) argued that descriptions of behavior can be developed that are context free, and these descriptions can then be interpreted alongside other emic information in the context of an organizational and a geographic setting. We discuss and expand on this conceptual framework later.
- That assuming that a test or instrument can be shown to offer equivalent psychometric functioning and a common metric for evaluating people's qualities irrespective of the language in which it is administered, and on the basis of the evidence that personality constructs such as the Big Five personality dimensions are generalizable across country, language, and cultural settings (D. P. Schmitt, Allik, McCrae & Benet-Martinez, 2007), predictors of performance can be constructed that are robust across multinational settings.
- That this combined emic–etic approach to developing a predictor–criterion model of the relationship between people's qualities and the organization's expectations for performance provides a theory-driven framework for evaluating the generalizability tests and assessments in a multinational context. It also provides a simple explanatory framework through which the value of what is being assessed by a test or assessment can be explained to stakeholders across the geographic locations of an organization who may not be testing or assessment specialists.

Another argument in support of this approach comes partly from organizational and cross-cultural research and partly from practical organizational need. The leadership literature has shown that leadership behaviors exist that are universally accepted as effective. For example, Bass (1997) and Bass, Burger, Doktor, and Barrett (1979) reported common qualities related to effective leadership across a range of diverse cultures and countries. House, Wright, and Aditya (1997) pointed to examples of how transformational leadership may be a common factor across culturally diverse settings, but the style in which transformational qualities are manifested also reflects different values within those settings. These findings suggest that it is possible to define commonalities in behavior and the personal qualities that relate to them while also capturing differences in the expression of behavior that reflect differences in the values within those settings. In other words, etic dimensions of behavior can be combined with emic dimensions of values to capture essential differences across contexts. The aim of the

combined emic–etic approach discussed here is very similar: to apply a generalizable framework that links behaviors to people's qualities, such as ability, personality, and motivation, while also capturing information about the style of behavior, such as values that reflect the organizational setting as well as the wider cultural setting in which work performance will be judged.

Practical organizational need is a critical factor. To paraphrase Weick (2001), organizations and those that lead and manage them are in the business of sensemaking, that is, making sense of the choices they have to make and of the consequences of the choices that are made. To enable organizations to make those choices and articulate them coherently, today's leaders and managers rely on metrics that enable them to set and measure achievements against objectives and targets across their organizations, and the management of talent is no different in its need for coherent and generalizable metrics. As Silzer and Church (2009) pointed out, defining what "good" looks like in terms of talent is no easy task, and it is dynamic because circumstances and organizational needs change. However, a criterion-centric framework based on a combined emic–etic approach is a significant contribution from the field to that endeavor, and one that provides a coherent strategy for developing solutions to the assessment needs of both national and multinational organizations. As an example, a multinational SJT developed for a global retailer is described next.

APPLYING THE COMBINED EMIC–ETIC APPROACH TO THE DEVELOPMENT OF SITUATIONAL JUDGMENT TESTS

The focus of the assignment was recruitment of customer assistants who would work in stores with responsibilities ranging from maintaining stock and assisting customers to checking out customers. Although the project was phased in with an initial proof of concept in the United Kingdom, the longer term multinational and multilanguage requirements were clear in that the solution would have to meet the needs of local operations across Asia, Europe including Turkey, and the United States. Within the United Kingdom, the client operated a variety of store formats ranging from smaller local stores to larger regional superstores, or "hypermarkets," so the test also had to be relevant to the working contexts across these different store formats.

SJTs are both an old and a new test format because items in the form of SJTs can be traced back to U.S. Civil Service exams in the 1870s and the work of Binet in the early 1900s (Weekley & Ployhart, 2006). Their more recent popularity can be traced to a publication by Motowidlo, Dunnette, and Carter (1990), and since that time journal articles and conference presentations on SJTs have grown almost exponentially (Burke, Vaughan, & Fix, 2009b). SJTs present several text, illustrated, animated or video-based scenarios and a set of possible response options. Respondents indicate which response they consider to be most effective or to most reflect how they would typically deal with the problem described in the scenario. As can be seen from the example provided in Figure 32.1, they appeal to job applicants and to stakeholders within organizations because of their immediate job relevance, and published research has lent them the reputation of being valid in terms of criterion-related evidence of validity (one aspect of Messick's external validity) and of being fair in showing small differences between reference groups (e.g., men, Whites, and younger job applicants) and focal groups (e.g., women, non-Whites, and older job applicants).

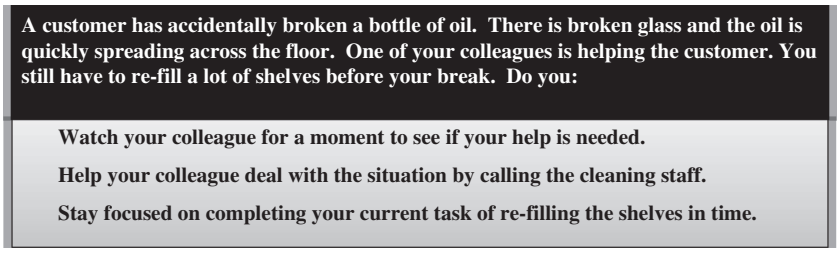
A customer has accidentally broken a bottle of oil. There is broken glass and the oil is quickly spreading across the floor. One of your colleagues is helping the customer. You still have to re-fill a lot of shelves before your break. Do you:
Watch your colleague for a moment to see if your help is needed.
Help your colleague deal with the situation by calling the cleaning staff.
Stay focused on completing your current task of re-filling the shelves in time.

FIGURE 32.1. Example of a situational judgment test item. Copyright SHL Group Ltd. Used with permission.

A specific need led to SJTs being the choice of assessment type: The job applicant might also be a customer, and so stakeholders within the organization had to be reassured that processes and assessments leading to a decision not to hire would also help to minimize the loss of a customer as a result of those hiring decisions. From the client's perspective, this need was seen as met if the assessment had high face validity and credibility to stakeholders and job applicants, and hence SJT became the assessment format of choice. To depart from the main thrust of this section of the chapter for a moment, the issue of the job applicant as an existing or potential customer has become a recent trend with clients, and it is one reason why SJTs are growing in popularity in Australia, New Zealand, the United Kingdom, the United States, and Western Europe, to mention but a few geographic areas in which this trend has developed.

SJTs tend to use one of three response formats: (a) rational, in which the judgments of subject matter experts are used to define scoring keys (e.g., which responses are more effective or indicate a better fit to a job or role); (b) behavioral tendency, in which respondents to an SJT item are asked to indicate which response would be most or least typical of them in a work situation; and (c) knowledge instructions, in which the respondent is asked to indicate which response would be more or less effective in resolving the situation portrayed in the SJT item. The term *knowledge instructions* is perhaps unfortunate in that it implies that the person taking the SJT has job-related knowledge that is being tested. In fact, the knowledge that is being tested is procedural and relational in nature rather than declarative, because the person's comprehension of the salient features of a situation and the responses most likely to lead to a resolution of that situation is what is being tested rather than any prior or acquired knowledge related to the specific job or role.

Meta-analyses synthesizing the results of several separate studies, such as those of McDaniel, Hartman, Whetzel, and Grubb (2007), have suggested that although knowledge instructions and behavioral tendency offer similar levels of criterion validity, the knowledge instructions format is more difficult to fake. Given that many of the assessment solutions used in employment settings are high stakes and are delivered online, we adopted the knowledge instructions format for this client.

Although SJTs have been used in personnel selection for several decades (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001), very little research has addressed how to best build and score SJTs (N. Schmitt & Chan, 2006; Weekley, Ployhart, & Holtz, 2006), and virtually no published research has addressed how to develop and evaluate SJTs in a multinational and multilanguage context. SJTs are also subject to a more fundamental question: Although many approaches have evolved for developing and scoring SJTs (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Weekley et al., 2006), what do they actually measure?

The practice of developing SJTs has largely been founded on an inductive and highly empirical approach, starting with critical incidents workshops (Flanagan, 1954) through which situations typifying effective and ineffective job performance are captured. The developers are then confronted with the task of inductively developing dimensions from the data gathered and then classifying both situations and the responses to them provided by subject matter experts (e.g., job incumbents and their supervisors or managers) into items and item clusters with scoring keys. Evidence supporting the validity of SJT scores tends to rely principally on criterion validation data showing the scores' relationship to performance ratings. Although criterion validity is critical to the evidence base supporting employment tests, it is only one aspect of the evidence proposed by Messick (1989, 1995, 1998) for external validity and, similar to empirical approaches to the development of biographical data or biodata instruments used to screen job applicants (such as the English method; see articles in Stokes, Mumford, & Owens, 1994; see also Chapter 25, this volume), empirical approaches to developing SJT scoring keys based on the relationship between responses to items and measures of job performance may have a very short half-life. This approach to developing a scoring key collects data on an SJT (or biodata questionnaire) from a sample of job incumbents and then identifies which responses to items are related to higher versus lower job performers. As such, the key may hold for that

sample of employees but may not generalize to a subsequent sample or to a future in which the dimensions through which job performance is evaluated have changed. The one substantive feature that this empirical approach lacks is a clear theory of the job or role and the qualities of people that are related to effective performance.

As such, one can claim that much of the practice in developing SJTs has lacked a coherent set of psychological constructs and a coherent measurement theory, and the deductive approach to these problems we developed seeks to address this gap by combining a clear set of constructs with several empirical checks to test the theory of the test, the scoring key for SJT items, and the relationship between SJT scores and job performance.

This deductive approach to SJTs also incorporates an interactionist model by using trait activation theory (TAT; Tett & Burnett, 2003) in the construction of SJT items. TAT focuses on the interaction between a person and a situation and seeks to explain behavior through the person's responses to trait-relevant cues within the situation. As such, the approach described here develops the situational component of an SJT item by inserting into that component cues or stimuli drawn from a taxonomy of traits within a meta-taxonomy of behaviors, as described later. The design of the alternates used to capture responses to the SJT's situational component measures the strength of the trait in terms of the answer options selected by the respondent. This fits well with a combined emic–etic approach in that the emic perspective is captured by workshops and through situations reflecting typical work situations in the context of a specific role or job in a specific organizational context. The etic perspective is captured through a defined set of trait constructs drawn from a context-free framework of work-related behaviors, which are used to index the work situations in terms of the traits related to them, thus providing emic–etic linkages used to construct assessment content. Readers may be interested to note that TAT has largely been researched and discussed in relation to assessment center exercises (e.g., Lievens, Chasteen, Day, & Christiansen, 2006) and what is described in relation to SJTs therefore has applicability to the development of assessment center exercises.

Our deductive approach to SJTs has been developed across several research and client engagements (Burke & Vaughan, 2011; Burke et al., 2009b) and uses the universal competency framework (UCF; Bartram, 2006; Burke, 2008b) to provide the behavioral and trait constructs through which a theory of the job or role is developed and through which TAT is applied to create a theory of the item. The UCF is the result of several years of research into the relationships between observed behaviors, as represented by competency models, and measures of personality, ability, and motivation (Bartram, 2005; Bartram & Brown, 2005; Burke, 2008b). The UCF operates at three levels of description, with eight factors, known as the Great Eight; 20 dimensions, with each dimension linked to a specific Great Eight factor; and 112 specific behaviors linked to specific UCF dimensions. Table 32.1 summarizes the first two levels of the UCF and shows how they link to personality as described by the Big Five motivation and cognitive ability constructs.

The UCF enables the articulation of an integrated theory of the job that brings together both predictor constructs (qualities of people such as cognitive abilities, personality, and motivational dimensions) and criterion constructs (the behavioral dimensions represented in the UCF). This facilitates a criterion-centric approach to specifying assessment requirements by first starting with the identification of those UCF behaviors that are critical to effective performance. For example, a key criterion for performance may be the effective management of projects, which would map to the UCF's Organizing and Executing factor and within that factor to the UCF's planning and organizing dimension. This dimension and more specific behaviors within it are known to correlate with facets of conscientiousness (a predictor construct), and therefore a theory of predictor–criterion relationships relevant to a specific job or role can be developed to guide the selection or development of tests and assessments (see Bartram, 2005, and Bartram, Brown, Fleck, Inceoglu, & Ward, 2006, for evidence of correlations between UCF constructs and other reference constructs such as the Big Five).

The language of the UCF is context free in description and can through a number of processes

TABLE 32.1

Overview of the Universal Competency Framework (UCF)

UCF factor	Example behavior (related UCF dimension)	Related predictor construct
Leading & Deciding	Makes prompt and clear decisions (deciding & initiating action); motivates & empowers others (leading & supervising)	Need for control
Supporting & Cooperating	Supports & cares for others (working with others); demonstrates integrity (adhering to principles & values)	Agreeableness, Need for Affiliation
Interacting & Presenting	Relates well to people (Relating & Networking); gains agreement & commitment from others (Persuading & Influencing); projects credibility (Presenting & Communicating Information)	Extraversion
Analyzing & Interpreting	Writes clearly & succinctly (writing & reporting); able to apply specialist knowledge (applying expertise & technology); makes rational judgments (analyzing)	Cognitive ability, Openness to Experience
Creating & Conceptualizing	Rapidly learns new tasks (learning & researching); produces new ideas and insights (creating & innovating); develops compelling visions for the future (formulating strategies & concepts)	Cognitive ability, Openness to Experience
Organizing & Executing	Manages time effectively (planning & organizing); focuses on customer needs & satisfaction (delivering results & meeting customer needs); follows procedures & policies (following instructions & procedures)	Conscientiousness
Adapting & Coping	Adapts to changing circumstances (adapting & coping); maintains a positive outlook (coping with pressure and setbacks)	Emotional stability
Enterprising & Performing	Accepts and tackles demanding goals (achieving personal work goals & objectives); identifies business opportunities (entrepreneurial & commercial thinking)	Conscientiousness, Need for Achievement

be mapped to the context of a job, role, team, or organization. In the example of the global retailer and the role of customer assistant, the mapping of UCF behaviors to the context of this role was executed through a series of workshops held initially in the United Kingdom with more than 60 experienced store managers participating and representing the full range of U.K. store formats. Each workshop ran for approximately 2 hours and first required each participant to review UCF behaviors and independently select and record on a standardized record sheet the six UCF behaviors they saw as being most critical to effective performance of customer assistants.

Once that task was completed, each workshop participant was asked to provide a short written description of a situation in which they had seen a customer assistant demonstrate effective performance. At this point, the deductive process was similar to the inductive process in that the critical incidents approach was used to capture context-specific information about work situations. However, once participants have recorded situations, in the deductive approach they are then asked to record which of the six UCF behaviors they previously selected as being critical to performance was best represented by the situation they described. The process was then repeated, with the participants being asked to describe a situation in which a customer assistant had not demonstrated effective performance and then to link that situation back to one of the six critical behaviors.

The data collected from these workshops thus provided a list of behaviors seen by organizational stakeholders as most critical to effective performance.

The consistency of the performance models held by stakeholders (as described by the six most critical behaviors selected by each stakeholder) could easily be evaluated by examining levels of interrater agreement. In the case of this client, performance models were found to be highly consistent and generalizable across store formats, and the most critical behaviors identified are summarized in Table 32.2. The data also provide a simple indexing of situations generated from the workshop by UCF behaviors because all situations can be tagged and organized by UCF behavior. This indexing process by UCF behaviors also provides the links to relevant predictor constructs, such as personality traits, through which TAT principles can then be applied to develop item blueprints for the construction of SJT items. For TAT to be applied effectively, the situation (a text description, video, or alternative presentation format such as a photographic storyboard or a graphical illustration) must contain a dilemma centered on a trait related to the UCF behavior of interest. For example, for a situation related to the UCF's following instructions and procedures dimension, the dilemma in the situation could be constructed around fulfilling an obligation such as completing a task versus abandoning a task to engage in another activity. Because this UCF dimension is linked to conscientiousness, knowledge of this trait can be used to construct a situation that acts as a stimulus to sample relevant facets of conscientiousness. In this way, the emic, or situationally specific, aspects of the item are captured while also ensuring that the item is developed within a clear and emic–etic predictor–criterion framework.

The second component of the item is represented by the response alternates. The more traditional approach to SJTs would tend to develop these alternates from the situationally specific information provided by critical incidents data and would rely heavily on the judgment of the analyst and the item developer. In the deductive approach, the biases of the analyst and the item developer can be mitigated by clear mapping of behaviors in the UCF to specific target traits and, where appropriate, shadow traits. For example, when a behavior has been identified that maps to perseverance on a task, a shadow trait could be applied that relates to variety seeking. Both traits can then be used to develop response alternates that operate as a scale ranging from the target trait through a neutral response to a shadow trait (when a shadow trait is not applied, then positive through negative examples of the target trait can be applied to develop the response alternates). Accordingly, the response alternates can be developed within an etic frame of reference to capture responses to the situation that are keyed to specific constructs and that provide a theory of the item that can be easily tested.

Once initial drafts of SJT items have been developed, the theory of the item is evaluated by asking

TABLE 32.2

Dimensions Identified for Customer Assistant

UCF factor	UCF dimension identified as most critical for customer assistant	Related predictor construct
Supporting & Cooperating	Supports & cares for others (working with others); demonstrates integrity (adhering to principles & values)	Agreeableness, need for affiliation
Interacting & Presenting	Relates well to people (relating & networking)	Extraversion
Organizing & Executing	Manages time effectively (planning & organizing); focuses on customer needs & satisfaction (delivering results & meeting customer needs); follows procedures & policies (following instructions & procedures)	Conscientiousness
Adapting & Coping	Adapts to changing circumstances (adapting & coping); maintains a positive outlook (coping with pressure and setbacks)	Emotional stability

Note. UCF = universal competency framework.

an independent panel of technical subject matter experts (e.g., behavioral psychologists or experienced item writers) to complete a simple rating exercise. Each panel member rates each alternate response to a situation on an effectiveness scale, and these data are then analyzed to see how consistent perceptions of effectiveness are across raters. In our experience, analysis of such data generated by six such subject matter experts is often sufficient to identify problems with items and scoring keys and to flag where the theory behind the item is weak or has failed, as indicated by low consistency across panel members on the scoring key for an item. This empirical check thus provides an early and cost-effective way to identify flawed item designs before larger scale field trials of items and statistical analyses of their characteristics.

The process just described can easily be extended, as it was in the case of the global retailer, to organizational stakeholders (in this case, store managers and team supervisors) across different geographic locations to provide data on the consistency of performance models across nationalities, languages, and organizational settings as well as to provide a pool of material from which generalizable item content can be developed. The idea of cultural decentering is discussed in more detail later in relation to the localization and adaptation of tests and assessments (see also Volume 3, Chapter 26, this handbook). Extending the process to a wider sample of stakeholders within an organization inherently enables decentering of the materials to be used in test and assessment development by avoiding those materials being originated in any one language or any one geographic context.

What evidence is there that the deductive approach yields useful measures? Given the constraints of space, two studies are presented. The first of these studies relates to the global retailer assignment and the evidence provided by four criterion-related validation studies conducted for the client and performance ratings of 342 employees. The second is a more recent study building on this client experience to test (a) the effectiveness of the deductive approach in constructing items that cluster into meaningful scales and exhibit the measurement properties generally expected of employment tests, such as internal consistency reliability, which SJTs generally fail to do, and (b) whether scores from SJTs using this approach correlate with the trait constructs they were designed to sample.

Consider again the global retailer assignment. The workshop process initially undertaken in the United Kingdom was repeated in the United States and in Turkey at the client's request. Item reviews conducted with organizational stakeholders in all three countries, coupled with analyses of item bias (methods for which are covered in more detail later in this chapter), resulted in 75% of all SJT items being common to all three versions of the SJT. Across language version pairs, 85% of the final content was common to the U.K. and U.S. English versions, 75% of the content was common to the U.K. English and Turkish versions, and 75% of the content was common to the U.S. English and Turkish versions. Although some content required adaptation, the level of adaptation was generally minor. An example is a situation developed from the original U.K. workshops that involved a bottle of wine falling from a shelf and breaking. The bottle of wine was changed to a bottle of oil to make the item suitable for use in the Turkish version, and the adapted version of the item will now replace the original in future language versions, showing the benefit of a multinational approach to instrument development. In general, then, the workshops identified a consistent set of critical behaviors from the UCF, and the deductive approach resulted in the major proportion of all content developed being deployed in all three versions of the SJT. The final versions of the SJT each contained 17 items and yielded a median internal consistency reliability of .76 across language versions.

Table 32.3 summarizes the criterion-related evidence obtained from validations of the customer assistant SJTs conducted in all three countries and languages. For the United Kingdom, two validations were undertaken, the first being a concurrent validation involving job incumbents who had not been selected using the SJT and the second being a predictive validation involving newly recruited customer assistants selected using the SJT. In the United States and Turkey, both validations were concurrent in design. To test the criterion relevance

TABLE 32.3

Criterion Validities for Customer Assistant Situation Judgment Test

Criterion measure	U.K. concurrent ($N = 78$)	U.K. predictive ($N = 70$)	Turkey concurrent ($N = 101$)	U.S. concurrent ($N = 93$)	Sample weighted average ($N = 342$)
Working with people	.35	.38	.31	.25	.32
Adhering to principles & values	.28	.18	.55	.21	.32
Relating and networking	.37	.52	.36	.29	.38
Meeting customer expectations	.46	.28	.22	.26	.30
Adapting to change	.38	.19	.26	.32	.29
Coping with pressure	.38	.18	.43	.10	.28
Overall job performance	.47	.40	.36	.28	.37

of the SJT scores against the critical UCF behaviors identified in the workshops, criterion measures were developed for supervisors and managers to rate employees on those UCF behaviors as well as on overall job performance. All criterion measures were administered in the raters' native language, and statistical checks (scale reliabilities and covariance structures) showed that criterion measures functioned equivalently across all three language versions. As shown in Table 32.3, criterion validities were consistent across all settings.

The practical value of these criterion validities can be appreciated by looking at Figure 32.2, which is taken from the original U.K. validation. The figure shows the proportions of those falling into the lowest and highest quartiles in terms of SJT scores that were also found to fall into the upper or lower quartile in terms of managers' ratings of performance. As can be seen, the odds of being a lower quartile performer for the lower quartile SJT score group were 5 to 1 (when the ratio of 63% divided by 13% is rounded to the nearest integer). In contrast, the odds of being an upper quartile performer were observed to be 12 to 1 for those in the upper quartile of SJT scores (46% divided by 4%). Overall, those in the upper quartile of SJT scores were 4 times more likely to be upper quartile performers than those in the lower quartile of SJT scores (46% divided by 13%). From feedback provided by the client from its own study conducted on the use of the SJT in staffing a new U.K. store, use of the SJT was estimated to have saved £1.2 million ($1.9 million US) by reducing the ratio of interviews to successful hires from 6:1 to 2:1. The applicants were also reported to like the assessment and to see it as a fair and credible method of identifying those who were more likely to enjoy a good fit between their talents and the skills required in the customer assistant role.

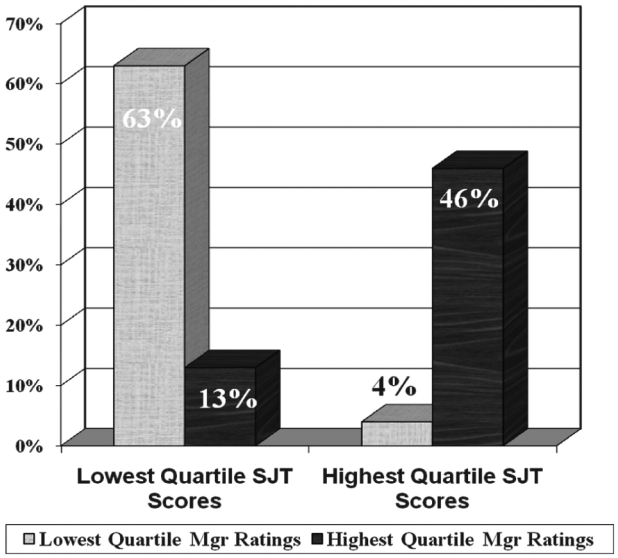

FIGURE 32.2. Comparative job performance by upper and lower quartile situational judgment test (SJT) scores. Mgr = manager. Copyright SHL Group Ltd. Used with permission.

Our experience with this assignment led to a research project to evaluate how effective the deductive approach was in capturing targeted constructs via an SJT format with meaningful measurement properties, again with customer service in mind.

Before sharing these data, we discuss a wider body of research that has explored trait constructs predicting performance in customer service roles. Hogan, Hogan, and Busch (1984) explored the attitudes and behaviors influencing the quality of the interactions between an organization and its customers or clients. Specifically, they looked at service orientation, which they saw as applying "to all jobs in which employees must represent their organization to the public and where smooth and cordial interactions are required" (pp. 170–171). Their research on various public sector positions showed that service orientation was most closely associated with what they referred to as likability (agreeableness) and adjustment (emotional stability) such that those seen as exhibiting a stronger service orientation were also more likely to be cooperative, rule following, and attentive to detail and to not be variety seekers; they were also more likely to be self-controlled, dependable, and well adjusted. More recently, Taylor, Pajo, Cheung, and Stringfield (2004) developed a scale for customer focus that they found to be related to three of the Big Five constructs—agreeableness, conscientiousness, and emotional stability. Digman (1997) also proposed these three constructs to be one of the higher order factors of personality that he has labeled *Alpha*, and which he defined as a socialization factor related to impulse restraint, conscience, and the management of hostility and aggression as well as neurotic defense. He made distinctions between agreeableness and hostility, conscientiousness and heedlessness, and emotional stability and neuroticism and proposed that those higher on Alpha are likely to exhibit higher impulse restraint and conscience. The reader will note that these three Big Five constructs map directly to most of the UCF behaviors identified as critical to the customer assistant role summarized in Table 32.2.

Using a bank of material obtained through application of the deductive approach to a number of client assignments and indexed to UCF behaviors, new SJT items were developed to represent three general clusters of situations related to customer service roles: situations and SJT items in which relationships with customers and colleagues were critical, labeled *people;* situations and SJT items in which the detail and organization of tasks were critical, labeled *tasks;* and situations and SJT items in which emotional self-management was critical, labeled *adaptability* (Burke, Vaughan, & Ablitt, 2011). In total, 25 items were developed. Figure 32.3 summarizes the results of an exploratory factor analysis of data obtained from a pilot sample of 210 participants who completed the SJT items and self-report items related to the Big Five personality dimensions. Each SJT item had three response options, as shown in Figure 32.1, and was scored 0, 1, or 2 in line with the TAT-driven scoring model for each item. To test whether the theory behind the item held, items were first analyzed within each situational cluster (people, tasks, or adaptability) to identify which items functioned most effectively within each respective cluster. As shown in Figure 32.3, four items were identified for each of the situational clusters on the basis of a single- or two-factor solution, and substantive weights for the items on the factors were identified within each situational cluster. This step of the analysis yielded 12 items in total across all three situational clusters, and these 12 items were then entered together into a maximum likelihood factor analysis assuming a correlated model. Although the goodness-of-fit chi-square indicates that the three-factor solution obtained for the 12 items does not account for all the covariance between items, and although one item had a low loading on one factor, the Kaiser–Meyer–Olkin statistic supports the appropriateness of applying factor analysis to these items (a value of 0.8 is generally accepted as indicating support for the appropriateness of factor analysis as a data analysis method), and the 12 items yielded an internal consistency reliability of .75.

For those experienced in psychometric and statistical analyses, these results may not appear remarkable, and the level of item attrition (12 of 25) does suggest that the process used to construct these items will benefit from refinement. However, taken in a context in which current wisdom states that SJTs are not amenable to factor analysis and tend to yield low internal consistency reliabilities, these results show promise for the deductive approach in providing a more adequate measurement model than traditional inductive SJT

People

Item No.	UCF Dimension
1	Working with People
3	Working with People
5	Working with People
10	Relating & Networking

Tasks

Item No.	UCF Dimension
13	Meeting Customer Expectations
14	Meeting Customer Expectations
15	Meeting Customer Expectations
16	Meeting Customer Expectations

Adaptability

Item No.	UCF Dimension
18	Adapting to Change
21	Adapting to Change
24	Coping with Setbacks
25	Coping with Setbacks

Pattern Matrix

Item No.	Factor 1	Factor 2	Factor 3
1	-0.12	0.33	0.41
3		0.65	-0.11
5	0.16	0.,34	
10		0.45	-0.12
13			0.65
14		0.16	
15		0.23	
16	0.11		0.32
18	0.37	0.16	0.14
21		0.69	0.10
24	0.96		
25		0.39	0.27

Summary Statistics

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.80
Goodness of Fit Chi -square Significance	0.05
Scale Internal Consistency (Cronbach's Alpha)	0.75

Factor Correlation Matrix

	Factor 2	Factor 3
Factor 1	0.52	0.40
Factor 2		0.43

Big 5 Correlations

	Factor 1	Factor 2	Factor 3
Agreeableness (A)	0.32	0.45	0.31
Conscientiousness (C)	0.28	0.45	0.24
Emotional Stability (ES)	0.14	0.21	0.17
A + C + ES	0.30	0.46	0.30

FIGURE 32.3. Evaluation of measurement properties and construct validity of a situational judgment test developed using the deductive approach. UCF = universal competency framework. Copyright SHL Group Ltd. Used with permission.

approaches. In terms of a psychological model and a theory of the job underpinning SJTs, the benefit of the deductive approach is supported by the correlations observed with Digman's (1997) Alpha personality constructs that the SJT items were designed to sample. Regressing the simple sum of SJT item scores (i.e., the total of 0, 1, and 2 scores across the 12 items) onto the reference scales for agreeableness, conscientiousness, and emotional stability yielded an *R* of .47 that, when adjusted for the unreliabilities of the scales in the analysis (the reference personality scales had an average internal consistency reliability of .84, and, as reported earlier, the SJT score had an internal consistency reliability of .75), the estimated operational (construct-level) correlation is .59.

Although SJTs are becoming more popular as a format of choice to meet assessment solutions, particularly with the opportunities to render this type of test as a more immersive assessment experience through technology (e.g., the use of video and dynamic avatars), they do present a challenge within any national or single-language setting by virtue of the question of what they measure. In the absence of a clear theory of the job and measurement models through which equivalence can be evaluated, challenges in a multinational setting need to be addressed with clear evidence available to users to assure them that SJTs deployed in a multinational setting will deliver assessment data on a common metric with consistent criterion validities.

Even when a combined emic–etic approach has been applied to developing and testing the generalizability of the constructs underpinning tests and assessments deployed in a multinational setting, the test or assessment may still need to be deployed in additional languages. Next, we consider the procedures and processes for the localization and adaptation of tests and assessments.

WHY LOCALIZATION IS MORE THAN JUST TRANSLATION AND EVIDENCE SUPPORTING A NEW APPROACH

Toward the end of 2006, Eugene Burke's employer made a request that, in turn, presented him with a problem. Having successfully developed a solution for secure online ability testing that had been rolled out in English, his employer wanted to know what investment and what time scale would be required to roll this solution out in a further 24 languages. A simple calculation informed Burke that the process would have to manage a volume of nearly 19,000 reasoning items in total.

The first step to try and resolve this problem was to consult experts in the field of localization and adaptation as well as to invite translation and localization service providers to submit proposals for supporting this program of work. The financial estimates received varied from the hundreds of thousands to several million, and the time estimates varied from several months to several years, hardly a consistent base from which to construct a budget and advise one's employer. The net result was, with Carly Vaughan's assistance, a review of the current wisdom in best practice for localization and adaptation of psychological tests and assessments and a deconstruction of that practice to develop a new process that delivered the program at a cost and time scale that was within the lower half of the various estimates obtained (Burke, 2009b; Burke, Bartram, Wright, Rebello, & Johannesson, 2008; Burke & Vaughan, 2010; Burke et al., 2009b). A critical factor in this process was to clearly state the roles and responsibilities of the various disciplines involved in localization (including project managers, psychologists, psychometricians, and translators) and to work backward from the metrics generally accepted as demonstrating construct and measurement equivalence across language versions to identify key risks and actions before the data collection stage of the process that mitigated against those risks. Given that this was where the efforts to develop localization and adaptation processes began, we briefly review the metrics for construct and measurement equivalence.

Geisinger (2003), drawing on Lonner (1979), pointed to four potential sources of bias when tests and assessments are localized from one language to another. *Linguistic bias* refers to issues that arise in the instructions and content of a test or assessment through wording, cultural references, idioms, and colloquialisms. *Conceptual bias* refers to whether the constructs underpinning the test or assessment can be generalized and calls for statistical evidence of *functional equivalence,* which is the third of Lonner's forms of bias. *Metric bias* refers to the lack of a single common metric such that comparisons across language versions are difficult or inappropriate to make.

Bartram et al. (2006) provided an example of statistical methods for demonstrating conceptual or construct equivalence. Using structural equation modeling (Bentler, 1990; Byrne, 2001) and data on 48,991 working adults across 13 languages (12 from Western Europe and the United States), they demonstrated that the pattern of scale intercorrelations from the Occupational Personality Questionnaire 32 were equivalent across the languages examined. Evidence of equivalence came from two structural equation modeling indices, the comparative fit index, in which values greater than 0.9 are considered representative of a well-fitting model, and the root-mean-square error of approximation, in which values less than or equal to 0.08 are taken to indicate a good fit of the model to the data. The median comparative fit index across language comparisons was found to be 0.982, whereas the median root-mean-square error of approximation was 0.019. In this case, the average sample size was 3,768, and although this exceeds the minimum sample size required to conduct structural equation modeling equivalence analyses, the reader will appreciate the time and cost that would be required to gather this level of data across the 24 languages as per the request to Burke.

Evidence of functional equivalence does not provide evidence of metric equivalence as addressed by methods for detecting differential item functioning (DIF). Hambleton, Swaminathan, and Rogers (1991) said, "An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (p. 110). DIF analysis serves to evaluate the extent to which items and scores on them place individuals from different groups on the same metric or whether the unit

of measurement used by an instrument is influenced by group membership. A useful methodology for evaluating DIF is that developed by Zumbo (1999) because it allows commonly available software such as Excel and the SPSS to be programmed to provide the analysis; the handbook developed by Zumbo for conducting DIF analyses is available as a free Internet download. When comparing different language versions of an item, this approach to DIF is essentially a hierarchical logistic regression of correct and incorrect answers to a test item first on an estimate of the person's trait (such as ability in the context of cognitive tests) and then on group membership and the interaction between ability and group membership. Where the model's fit improves substantially with the addition of group membership, evidence of uniform DIF is provided such that the item has similar function for the groups compared, but the groups have different zero points on the scale of the trait examined. When the model's fit increases substantially with the introduction of the interaction between the trait and group membership, then evidence is provided such that the item functions very differently for both groups. Although DIF has most commonly been applied to measures of cognitive ability and educational attainment, it can be applied to noncognitive measures, and Burke et al. (2010) provided examples of how DIF was applied to evaluating the language equivalence of a forced-choice questionnaire designed to screen for counterproductive behaviors.

In conducting DIF analyses for both cognitive and noncognitive items, our experience has suggested that a minimum sample of 150 for each of the two groups compared provides little loss in the power to detect biased items and that a ratio of 2:1 between the reference group (e.g., those tested in the original English version) and the focal group (those tested in a non-English version) does not introduce undue statistical bias into DIF analyses. This recommendation is based on a combination of empirical and simulation studies. However, and in the context of the request put to Eugene Burke in 2006, rolling out 24 other language versions from an original English-language version would amount to gathering data on a minimum of 3,750 trial participants. As such, any issues identified at this stage of the localization process would prove costly to remedy and would require repeating the data collection and analysis processes.

From our review of various guidelines and tests related to localization and adaptation (e.g., Hambleton, Merenda, & Spielberger, 2005; International Test Commission, 2010), and determining the risks that could result in a failure to obtain construct or measurement equivalence, the most obvious place to start in designing a localization process is keeping the most obvious risks in mind. Figure 32.4 provides an example of the numerical and verbal reasoning test items to be localized, which are designed to reflect common tasks and contexts in contemporary working environments. As such, the extent of content bias

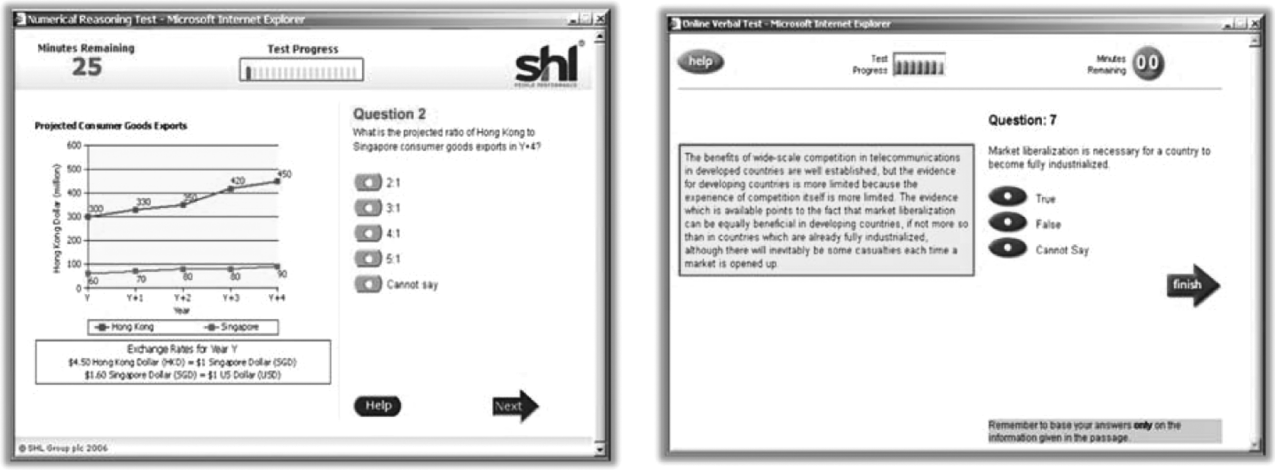
Numerical Reasoning Test - Microsoft Internet Explorer
Minutes Remaining
25
Test Progress
shl
Projected Consumer Goods Exports
Hong Kong Dollar (million)
Year
Hong Kong
Singapore
Exchange Rates for Year Y
$4.50 Hong Kong Dollar (HKD) = $1 Singapore Dollar (SGD)
$1.60 Singapore Dollar (SGD) = $1 US Dollar (USD)
Question 2
What is the projected ratio of Hong Kong to Singapore consumer goods exports in Y+4?
2:1
3:1
4:1
5:1
Cannot say
Help
Next
© SHL Group plc 2006
Online Verbal Test - Microsoft Internet Explorer
help
Test Progress
Minutes Remaining
00
Question: 7
The benefits of wide-scale competition in telecommunications in developed countries are well established, but the evidence for developing countries is more limited because the experience of competition itself is more limited. The evidence which is available points to the fact that market liberalization can be equally beneficial in developing countries, if not more so than in countries which are already fully industrialized, although there will inevitably be some casualties each time a market is opened up.
Market liberalization is necessary for a country to become fully industrialized.
True
False
Cannot Say
finish
Remember to base your answers only on the information given in the passage.

FIGURE 32.4. Example Verify items. Copyright SHL Group Ltd. Used with permission.

(that the items would be sampling content domains that are not generalizable across languages and geographic areas; see Geisinger, 2003, for examples) was felt to be low. However, linguistic and cultural biases and—because these tests were to be deployed via the Internet and therefore displayed on a wide range of screen formats at the respondent's physical location—method biases such as screen formatting were seen as essential factors to evaluate first. The first step in the process was to conduct an initial localization risk review to identify these factors upfront.

One example item contained the question "The ratio of sales of conservatories compared to PVC window installations changed by what ratio between Year 1 and Year 3?" Although this item would have relevance to a U.K. audience, it would probably have little relevance to a U.S. audience and to audiences in other languages (this was picked up on an initial review for U.S. English). One challenge was to determine whether different language versions of the test would be required for deployment in South American Spanish-speaking countries. Consider an item that refers to a department store. This term varies across South American languages as well as in the Spanish spoken in the United States, but the word *store* is common. As such, the removal of the word *department,* which is redundant to the item's functioning, enables the item to have wider language availability. Trivial as these examples may seem, their impact can be significant.

To conduct the localization risk review, and indeed the full localization process, a key requirement was to ensure that the process was staffed with personnel with the right skills and that work roles and responsibilities were clear. As such, the review and the overall process were managed by a testing specialist working with a translation project manager and two experienced and qualified translators. The requirement for translators was that, in addition to the appropriate technical qualifications, they have at least 5 years experience translating from the target language (e.g., Hungarian, Indonesian, Norwegian, or Japanese) into English and that they be situated in the country in which the target language originated. The latter requirement was made to ensure that translators were current with the contemporary use of the target language.

To participate in the review and subsequent localization process, translators were required to participate in an orientation session delivered via webinar that explained the purpose of the items and tests and also explained the structure of the items and how item stems (e.g., text passages or numerical tables and graphs) related to response alternates. As such, the task and contribution expected of translators versus those expected of testing specialists were distinguished and specified, with a clear focus on the specific tasks and responsibilities of each stakeholder in the process. This process differs somewhat from several guidelines that suggest that linguists and translators undergo an introductory course in tests and measurements, and this source of expertise was provided throughout the process by the testing specialist project manager. Translators were then required to undertake a test translation to check their understanding of the principles and tasks involved, which then determined whether they were accepted into the program or replaced by an alternate translator.

Before we describe the localization process in more detail, note that before the localization process, all items had been coded by the test development team under Carly Vaughan's guidance. This coding identified critical content in the item stem, the response alternates, and those items that required as literal a translation as possible to manage any linguistic or other biases affecting the items' functioning in other languages. The translators' training explained the coding in relation to an item's functioning (i.e., the thought process linking the content of the item to correct and alternate item response options) and when near-literal translation was required versus when more license for the translator was allowable. The test translation tested the translator's understanding of this requirement.

The localization review and subsequent fuller localization process involved two translators per language. Content was divided between the two translators, who used a standardized record sheet to record any problems they foresaw in terms of words, phrases, sentence structure, and other item content elements. Through the translation project manager, these independent reviews were then collated and brought to a web conference in which all issues were

discussed among the testing specialist project manager, the translation project manager, and the two translators. The output of this conference was agreement on specifications for the subsequent translation of items, including any amendments to the original items, thereby capturing and recording any adaptations required in the original English version (effectively a decentering of test content).

Recall that the items were intended for deployment via the Internet and that formatting issues were identified as a potential method factor to be evaluated. Guidance from human factors colleagues provided screen areas within which item content had to be maintained to meet minimum screen sizes for which the online test was designed. Part of the localization risk review served to identify formatting issues, such as text length increasing through translation to a target language, and whether alternate font sizes could be used to resolve the problem or whether adaptation of the text was required to meet the screen format standards and enable legibility when deployed in the target language.

As mentioned earlier, each translator was responsible for translating half of the test content, with the content being exchanged between translators after each translator had completed his or her translations; each translator then reviewed his or her colleague's work, comparing the translated content to the English original. Comments were then collated by the translation project manager and brought to a harmonization meeting chaired by the testing specialist project manager (the title of this meeting was chosen by all parties involved to promote the idea that any issues needed to be resolved in a collegiate and objective way). This point in the process allowed for any final issues to be resolved and for decisions on the final translations before trials of the content with native language speakers.

Those familiar with the literature on localization will note that this process follows principles such as decentering, in which content is adapted in the original language to improve its translatability as well as elements of the committee approach, with multiple host and target language speakers involved. The one element of common practice whose absence the reader experienced in localization and adaptation will notice is back-translation. As described by van de Vijver and Leung (1997), this practice involves an initial translation from the host to the target language followed by an independent back-translation from the target to the host language. Here are some examples from earlier localization efforts with noncognitive tests and assessments that explain why a back-translation was not pursued: (a) The original text asked respondents to rate how often they get angry or upset, and the back-translation resulted in the text *gets angry quickly or upset easily,* making the item less attractive, and (b) *is motivated to do well in their job* became *gets involved to do his job properly,* making the sentence more complex and changing it from a statement about achievement to one more strongly framed as compliance.

By structuring tasks and responsibilities as well as respecting the technical and professional contribution of the specialists involved, and by directly evaluating linguistic, cultural, and method risks at the earliest stages of the localization process, the added complications of translation and back-translation are avoided, and the risk of meaning being lost in translation is minimized. As evidence supporting this statement, Figure 32.5 shows the results of DIF analyses conducted on a sample of the languages into which the reasoning items were localized. The figure compares DIF rates against an earlier process using translation and back-translation. Note that Finnish was a new language,

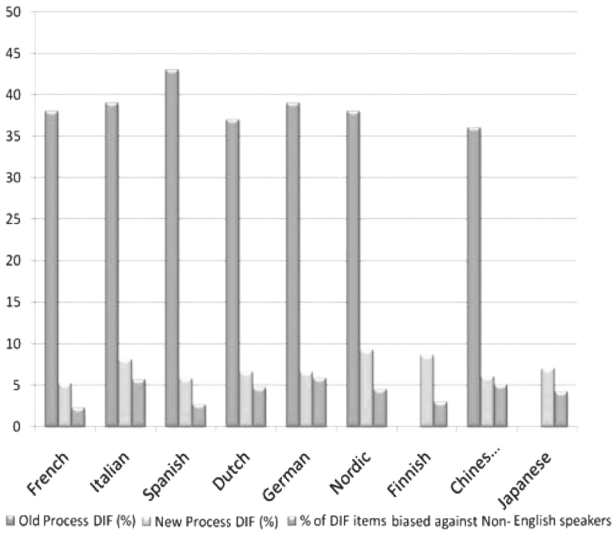

FIGURE 32.5. Differential item functioning (DIF) rate.

and therefore comparative data on the older process were not available, but the reader will also note that the DIF rates obtained from the newer process for this unique language are comparable to those for the other languages shown.

Thus far, we have shared experiences and processes to ensure that tests and assessments measure meaningful constructs that are generalizable across multinational settings and that provide information from which comparisons of people can be made on the same metric irrespective of their country of origin or their native language. The next issue is the norms against which scores are compared and that play a significant part in how test and assessment scores are used in the decisions made about employment opportunities.

LOCAL OR GLOBAL: WHAT NORM IS APPROPRIATE IN MULTINATIONAL SETTINGS?

Thus far, our focus has been on whether differences in scores observed for tests and assessments in multinational settings may be a factor of lack of equivalence across language versions of the test or assessment. Now we will consider the issue of whether differences in scores are a factor of the people and what the nature of those differences mean in multinational settings.

Norms (whether they are percentiles, *T* scores, stens, stanines, grades, or some other form of normative score) are used to provide information on a person's relative standing in comparison with others. In discussing the relatively new territory of international or global norms, consider the question of whether it is the people. The online *Oxford English Dictionary* defines a *norm* as "the usual or standard thing" and as "a required or acceptable standard." It also defines a norm as "a general rule regulating behavior or thought," and in employment testing, the accepted practice has been to provide local or user norms or, in the case of personality tests, the aggregation of user norms at the country and language level to create local national and language norms (Bartram, 2009).

So how do multinational users of tests and assessments select the appropriate norm? Do they operate within the constraints of nationally or language-defined norms and use a norm for each version of the test or assessment, or do they operate within each geographic area? If they do, how will they know that the normed data provided by tests and assessments is consistent in what it communicates to them about their talent needs? How will they know whether any significant factor has been overlooked in terms of a meaningful and substantiated difference in talent across the organization or whether the norms they are using mask a bias that may lead to ill-informed interpretations and decisions?

What drives the norm is a question that Roe (2009) set out to address by requiring that those involved in the provision of norms state which factors influence scores and, therefore, which factors the test score user should consider in selecting an appropriate norm. Roe's conceptualization assumes that the generalizability and measurement equivalence of the test and assessment scores has been determined, from which assumption the investigation can move on to several factors operating to influence scores and, therefore, norms. The framework Roe proposed encompasses endogenous factors such as gender, age, and ethnicity in the sense that, if these are factors influencing test scores, then their representation in a norm group will be important in the understanding and relevance of a norm. Roe also proposed exogenous factors such as educational level and type and job level and type as well as industry sector and organizational type, to which can be added nationality and language. The third set of factors he labeled examination factors, which include test format (e.g., computer or paper and pencil) and whether the test is administered in a high- or low-stakes setting, and to which can be added (as described later) whether the test is administered in a proctored or unproctored environment. The final factor in Roe's framework is that of time, which refers to the time frame covered by the norm as well as generational factors within any one cross-section of a population at a specific time that may influence scores (see Flynn, 2007, for a full and intriguing discussion of how generational differences can be misinterpreted and how they may represent true changes over time in reasoning abilities).

Tett et al. (2009) reviewed a number of personality test manuals and commented that given that several norms were provided for single tests, the choice of which norm is appropriate is often difficult for the user. Noting the work of Ang, van Dyne, and

Koh (2006); Judge and Cable (1997); and Warr and Pearce (2004) showing that personality scores are related to job type and to preferences for organizational culture, relationships that one might expect from Schneider's (1987) attraction–selection–attrition mode, Tett et al. called for norms that offer greater clarity in terms of relevance to the user and that reflect job type and organizational characteristics. However, these authors proposed greater use of local norms that, although it may be effective in some national and organizational settings, may leave the test or assessment user short on important information in multinational settings.

Consider the example of a multinational bank that is a strong advocate for testing and assessment in all its operations worldwide and that seeks to maintain a consistent policy and standards in its recruitment and selection of staff. Local norms might serve to guide local operations, but consider how such norms would serve this organization in comparing talent across as well as within geographic areas by business function and job level.

Wright (2009) reported the results of investigating the relationship between cognitive ability test scores and two of Roe's (2009) exogenous factors, industry sector and job level (the latter being strongly related to the educational levels within the populations of job applicants examined). The data were drawn from a number of U.K. local client norms that had been created for verbal and numerical reasoning tests sharing the same underlying constructs and item formats and were designed to operate at managerial and professional job levels through semiskilled operational roles. The norms covered a variety of industry sectors including banking, financial services, professional services, science and technology, manufacturing, retail, leisure, and local and national government. For 95 groups and a total sample of 52,300 job applicants, test score means were found to cluster around four major industry sector groupings: (a) banking, financial, and professional services; (b) science, technology, and manufacturing; (c) retail and leisure; and (d) public sector (local and national government). Standard deviations were found to reflect factors such as the educational mix of populations, as indicated in Figure 32.6.

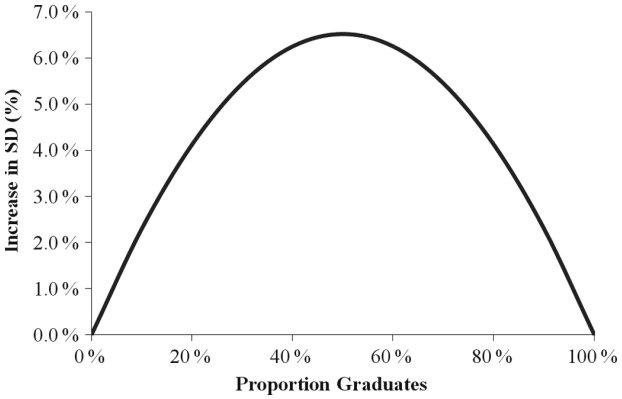

FIGURE 32.6. Relationship between score standard deviations and proportion of graduates in the norm sample.

Burke (2009a) developed Roe's (2009) conceptual framework, using the statistical evidence provided by Wright (2009) to suggest that norms should be conceived and developed from the perspective of talent metrics that enable organizations to answer questions such as how effective their processes have been in attracting and acquiring talent relative to other organizations in their industry sector. This approach to norm development was applied in the construction of comparison groups for the Verify solution to unproctored Internet testing (UIT; Burke, van Someren, & Tatham, 2006) and offers a norm framework with several job levels by four industry sectors (as per Wright's findings) and a general population reference group. These comparison groups also reflect endogenous and exogenous factors that typify the demographics, educational levels, and job types found among applicants to organizations. Although this framework was originally developed with U.K. data, more recent analyses have shown it to generalize to a broader international context. An example is the analysis of numerical reasoning scores obtained from 8,432 in vivo administrations of Verify to job applicants in Belgium, Denmark, France, Germany, Italy, Norway, Sweden, the Netherlands, the United Kingdom, and the United States, which showed that scores are influenced by educational attainment (a key distinction being the attainment of a high school degree, baccalaureate, or higher educational qualification), industry sector (in line with the findings from Wright and the industry sectors in the Verify

comparison group structure), and the business function for which the person had applied (a key distinction being applicants for financial, professional, and research functions vs. other business functions). The relationships identified by Wright were found to hold irrespective of the applicant's nationality.

Analysis of data from 337,646 in vivo administrations of the Occupational Personality Questionnaire 32 across 19 different countries showed that the careful aggregation of data sensitive to the weighting of exogenous factors such as gender and endogenous factors such as country could be used to develop an international norm for a personality instrument (Burke, Bartram, & Philpott, 2009). In line with findings of differences by country reported by Bartram et al. (2006) in the Occupational Personality Questionnaire 32 technical manual, comparisons of country profiles using this international norm and the score profile within each country showed that most differences were small to medium across most scales compared. Authors of this norm noted the differences identified by country and language and proposed guidance for the user in choosing whether to use the international norm or whether to use a local national or language norm by considering the consequences of the choices made by a potential user. The key point here is that users are presented with a clear choice guided by knowledge of the decisions for which they want to use assessment data and explicit information on the endogenous and exogenous data related to that choice.

Returning briefly to the international bank, at the heart of its question were concerns over how well the bank was doing against the competition for talent in its sector and across its geographic areas and business functions and how consistent the bank was in acquiring the quality of talent seen as essential to its organizational objectives. So, the bank's fundamental need was for what can be called a *talent mark,* or a means of understanding how the talent it has acquired measures up when compared with its industry sector globally. Data on verbal and numerical reasoning test scores for 1,173 employees recruited by this organization showed a good fit to the banking, financial services, and professional services comparison group described earlier. The data covered four geographic regions (the Americas, Australia and New Zealand, Asia and Europe, and the Middle East and Southern Asia), seven job levels, and nine business functions. Little variation was found by region or by job level, but meaningful differences were found by business function, with some exceeding the industry benchmark and others falling slightly short of it. We hope the reader will see how a norm in the form of a talent mark can enable organizations such as this one to understand how effective their talent processes are in achieving organizational objectives and how such data help organizations to consider their future talent strategy.

Building on Roe's (2009) conceptual framework and the applications just described, one can conceive a hierarchical framework of norms for both can-do and will-do tests and assessments. Such a hierarchy could operate at several geographic levels, from the global level through the regional and national levels as well as encompass industry sectors, business functions, and job types and levels. The key to such a framework or taxonomy of norms is to address the issues highlighted by Roe, who put forth the following challenge to test and assessment providers:

> In this approach the test developer is challenged to decide which factors and interactions must be considered and which ones can be safely ignored before thinking about potential reference groups. . . . The choice of a reference group follows primarily from the purpose of test use, i.e. the type of comparisons needed to provide clients with meaningful information and allow them to take unbiased decisions. . . . This underlines the need for greater awareness of contingency factors and the importance of collecting relevant data in the future.

GOING ONLINE WITH TESTS AND ASSESSMENTS AND THE ISSUE OF SECURITY

One of the more controversial developments in employment testing in the early 21st century is that of UIT. The strong opinions on this topic were

exemplified in the article by Tippins et al. (2006). Thinking and research on this topic have moved on since this article and, in particular, since the Society of Industrial and Organizational Psychology conference symposium organized by Tippins in 2008 (see, e.g., the special issue of *Industrial and Organizational Psychology* published in 2009 on this topic). Positions have also shifted, as evidenced by articles such as that by Arthur, Glaze, Villado, and Taylor (2010), who found few differences in proctored and unproctored ability test scores, although the authors caveated their findings by pointing out that the type of test investigated might be less susceptible to the effects of cheating. Drasgow, Nye, Guo, and Tay (2009) have even suggested that the supposed gold standard of secure testing, proctored test administrations, is, in reality, a misnomer, and they suggested that the belief that the presence of a proctor somehow guarantees security is a false assumption.

Before describing one solution to the problems of UIT, we first address the issue of whether there is a real threat to the security of UITs. Although researchers such as Arthur et al. (2010) have suggested that there may not be, Tate and Hughes (2007) reported results from a survey of 319 university undergraduates' and postgraduates' perceptions of UIT across 51 U.K. universities. The vast majority (76%) had taken UITs at home, and 81% of respondents reported this administration option as the most preferred. Asked to report the frequency of actions that were inappropriate while taking UITs, about one-eighth of respondents (37, or 12%) reported actions that could be constituted as cheating, and among those respondents some reported colluding with friends, obtaining the questions in advance, and circumventing the technology in some way. When respondents were asked what would deter them from cheating, the top response was their own honesty (77%) followed by seeing no long-term advantage (47%) and fear of being caught (35%). Our view is that the threat to the security of tests in general and in any high-stakes setting and for UITs is real.

Many of the solutions to this problem are variants of Seagall's (as described in Tippins et al., 2006) two-step verification process, and we explore one such solution developed through the Verify program (Bartram & Burke, in press; Burke, 2006, 2008a, 2008c, 2008d, 2009a; Burke, Mahoney-Phillips, Bowler, & Downey, 2011; Burke et al., 2006; Lievens & Burke, 2011). In this solution, an initial UIT is administered, followed by a subsequent and proctored verification test administered to those proceeding to a later stage of, say, a recruitment process. This approach extends that suggested by Seagall to encompass Impara and Foster's (2006) principles for the security of testing programs by including features aimed at defenses against cheating before test administration as well as during test administration. The latter is the principal focus of Seagall's proposal. One such proactive action is to conduct web patrols looking for evidence of content piracy and for the unauthorized exchange of materials. One such web patrol captured the thread shared earlier. Indeed, the Internet, seen by some as the Achilles heel of UIT, actually provides a key element in defending the security of employment tests (whether administered by UIT or not) because it is also a key medium through which those seeking to undermine test security carry out transactions and advertise materials.

The following principles, developed as a contribution to the Association of Test Publishers Test Security Summit, serve as a summary of the key objectives of the Verify program as well as its key features (Burke, 2008c):

- enforcing test security by actively managing intellectual property breach and monitoring candidate behaviors;
- identifying test fraud by policing content and monitoring through critical incident procedures and regular data audits to check for piracy, cheating, and item exposure; and
- preventing test fraud by designing cheat resistance into the score of record.

Figure 32.7 summarizes how Verify designs security into test administration. The first UIT administration draws on multiple test forms using the linear-on-the-fly model (Davey & Nering, 2002). Drawing on item banks calibrated through item response theory, the system continually creates equivalent test forms that are first checks for psychometric quality and, if they meet those checks, are then

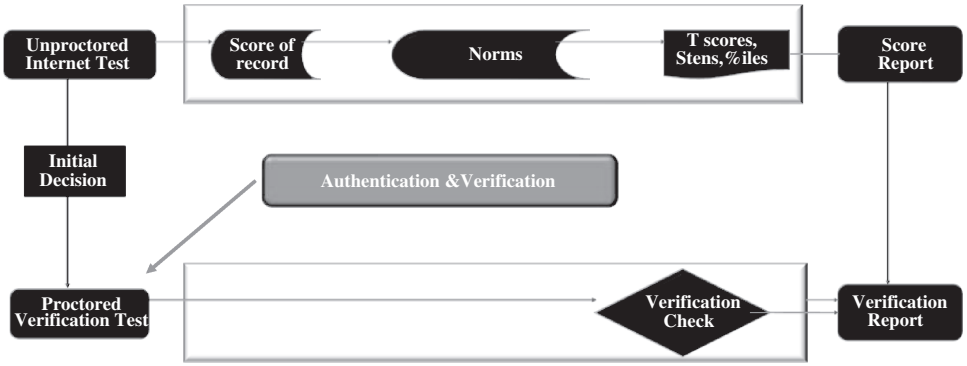

FIGURE 32.7. Conceptual overview of Verify.

registered in a test bank. When a candidate registers for a test (this process is compliant with the recommendations of the International Test Commission, 2005), one of the forms held in the test bank is then randomly assigned to the candidate. Once the candidate completes the test, all information is returned to a secure server where answer keys are held and that are not downloaded to the candidate's computer. Accordingly, item exposure and the security of answer keys are maintained. Before taking the test, the candidate is asked to agree to an honesty contract, playing to the deterrent noted by Tate and Hughes (2007), and are informed that should they proceed through the process, their score will be subjected to a further verification check obtained through the administration of a proctored verification test at some later point. Note that the score of record (i.e., the score used in decision making) is the score from the initial UIT because the verification test serves purely to provide data to validate the first UIT score.

In addition to these within-administration security features and as shown in Figure 32.8, the process is supported by a wider security framework of web patrols, critical incidents reporting, and data forensic audits (see Maynes, 2009, for further details on these statistical audits). As reported by Burke, Mahoney-Phillips, et al. (2011), data forensic analyses of the first-stage UITs showed low frequencies of abnormal and aberrant question responses and overall test scores (e.g., 0.003% of 30,000 test administrations were found to display fast latencies and high question accuracy, where faster and more accurate responses may suggest prior access to the test's answer key). Overall, data forensic analyses of Verify data have shown that 2% of applicants have one or more data forensic indices flagged as abnormal or aberrant. Although low, 2% of 100,000 applicants (typical of some recruitment programs) could suggest that 2,000 applicants might achieve scores exceeding the cut-score levels set for various client testing programs, emphasizing the need for additional security measures such as verification testing. Data on the frequency of inconsistent scores observed from the administration of verification tests have indicated that this threat varies depending on the nature of the testing program and the type of candidate. Tate and Hughes (2007) estimated that the base rate for cheating in European graduate recruitment testing programs is about 12%, or 1 in 8 candidates. However, data for graduate (campus) recruiting campaigns have shown rates that are 2 to almost 3 times higher than would be expected by chance alone and are comparable to the estimate for cheating given by Tate and Hughes. It is important to note that nonverification does not, in itself, demonstrate cheating because other factors such as the change in administration from UIT to a proctored setting, the person's health on the day of the test, and the person's emotional state may all affect test scores, but such findings do suggest that these highly competitive and high-stakes campaigns are those for which the need for greater security in employment testing is likely to be paramount.

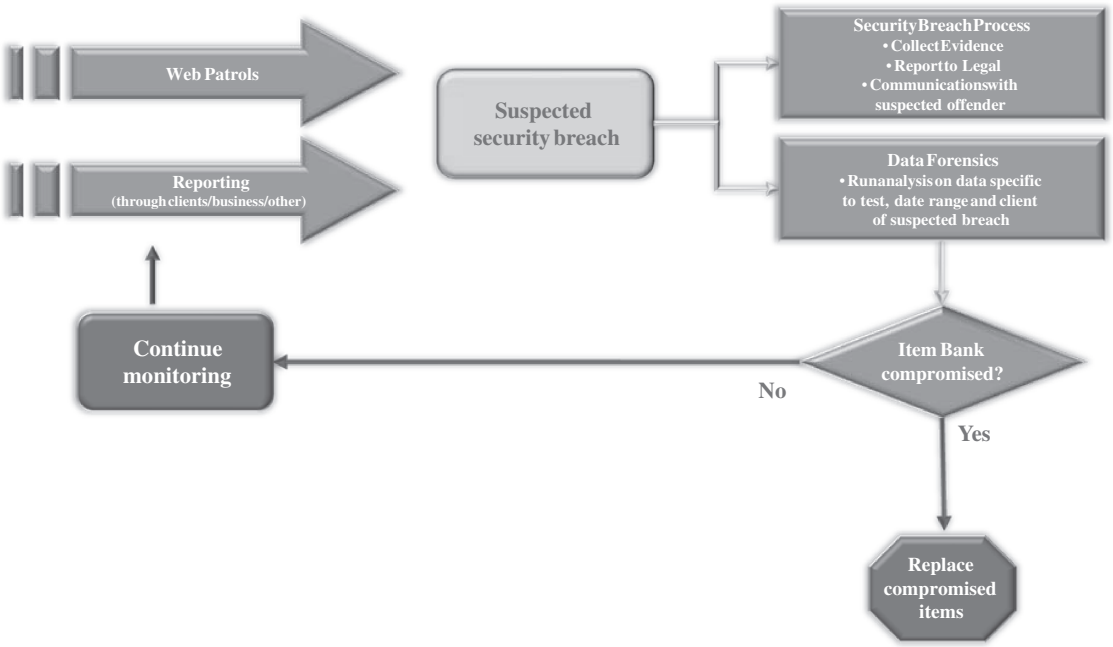

FIGURE 32.8. Verify security framework.

A final note on the issue of test security: It is truly a global issue. Web patrols conducted since 2006 have found pirate sites in the United Kingdom as well as China, and of particular note are pirate sites in China offering content in English, reflecting the importation by Western organizations of recruitment practices such as employment testing and assessment into other geographic locales (Burke, 2008b). Test security is an issue but, as a science and as a practice, the understanding of attitudes toward cheating on employment tests has grown substantially in the past few years. Solutions such as that described do go a long way toward deterring cheating and piracy, and technology in the form of test delivery systems (such as linear-on-the-fly), methods for detecting abnormal score patterns, and more sophisticated methods of scoring offer substantial advantages in delivering cost-effective assessment that, in turn, delivers value to organizations in their acquisition and development of talent.

WHAT OF THE FUTURE OF EMPLOYMENT TESTING IN MULTINATIONAL SETTINGS?

One trend that is easy to predict from the invitation to contribute this chapter is that multinational employment testing and assessment will continue to spread as globalization and the migration of labor continues. Indeed, new challenges will emerge as the nature of work becomes more diverse with people from different national, language, and cultural backgrounds interacting as members of physically colocated teams and as communication technologies provide easier, faster, and more accessible means of communicating. So, new competencies will grow in importance, as evidenced by the GLOBE project and the need to meet the demands of international and global work.

Testing and assessment have always been shaped and influenced by technology. Today, technologies exist that were not accessible 10 or 15 years ago, with simulated worlds, avatars, and social networking sites growing in popularity as well as video communications, smartphones, and tablet technologies. Animated versions of SJTs already exist, but the challenges outlined in this chapter will need to be addressed if the packaging of more immersive forms of assessment environments is to live up to the standards society has come to expect of valid assessments and these assessments are to truly meet the needs of

organizations. Whatever the measurement challenges might be, it would be folly for science and practice in assessment to ignore the need to exploit these new technologies, and candidates' demands for an assessment experience that is comfortable for them is one strong force that will act to strengthen the impact of technology on assessment.

One trend that has already emerged and will challenge many accepted notions of the technical qualities of tests and assessments is the demand for ever more efficient forms of assessment. With the growth of the Internet as the preferred means of delivery, the notion of long tests will be challenged. Burke and Bateson (2009) commented on this in describing how the criterion-centric perspective offers the opportunity to move from elaboration of predictors in the form of fairly long personality and ability tests to shorter, multicomponent forms of assessment in which the focus of design is on composite scores that capture predictor–criterion relationships rather than on the more traditional profile sheets that tend to talk to the person rather than to the fit of the person to a job, role, or organization. Burke, Mahoney-Phillips, et al. (2011) described one application of this approach to meeting the needs of an international bank whose key requirement was that the fit of an applicant to one of three roles had to be delivered in 30 minutes or less. Solutions such as this challenge classical notions of singular scales and internal consistency reliabilities.

As much as these demands present challenges, they also present opportunities, such as the development of new measurement models. One example is the development of item response theory models for forced-choice self-report questionnaires (Bartram & Burke, in press). Although these models were originally developed to provide an effective measurement model for the delivery of fakeproof self-report measures, they offer potential to develop shorter and much more efficient assessments with little loss in the fidelity of those assessments. As another example of this trend, Burke et al. (2010) described the validity of a short screening questionnaire for counterproductive behaviors in the format of a criterion-oriented personality scale that takes only a matter of minutes to administer and that can be easily bundled as one component of an assessment solution. In the space devoted to SJTs, we have suggested many of the principles for assessing constructs through more interactive and situationally or simulation-based approaches, and extending these notions to more efficient forms of assessment requires only a short conceptual step.

How scores are reported and used as talent analytics is another likely growth area. Traditionally, reports have tended to center on the person and on his or her qualities, with an emphasis on how reliably those qualities have been estimated. In the future, reports may possibly provide actuarial information in terms of future performance and longer term potential and, through technology, enable the user to determine in more detail choices in terms of the actions they can take to leverage potential at the individual, team, and organizational levels. Organizations have a voracious appetite for metrics, and talent management is no less hungry a client.

Finally, in terms of crystal ball gazing, is a move from user-centric models to more candidate-centric ones. To date, testing and assessment solutions have tended to provide data on an individual in response to an organization's needs to consider that person in the context of a job or role at a single point in time. The customer in this model is the organization. Consider an alternative model in which the candidate provides data in terms of his or her talents and potential in search of a role and an organization that best fits those talents and potential. In such a model, the data center on the candidate rather than on any one organization, and the assessment data provide the opportunity for the candidate or agent to search many potential clients for that person's talents.

Perhaps the best prediction for the future of multinational testing and assessment is that it will provide challenges but also opportunities for innovation because, in the words of the computer scientist Alan Kay, "The best way to predict the future is to invent it."

References

Ang, S., van Dyne, L., & Koh, C. (2006). Personality correlates of the four-factor model of cultural intelligence. *Group and Organization Management, 31*, 100–123. doi:10.1177/1059601105275267

Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and

response distortion effects on unproctored Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*, 1–16. doi:10.1111/j.1468-2389.2010.00476.x

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203. doi:10.1037/0021-9010.90.6.1185

Bartram, D. (2006). *The SHL universal competency framework.* Thames Ditton, England: SHL Group.

Bartram, D. (2009, July). Why do we use norms? In E. Burke (Chair), *Reconsidering norms: Their construction and their value in enabling valid decisions*. Symposium held at the European Congress of Psychology, Oslo, Norway.

Bartram, D., & Brown, A. (2005). *Great Eight factor model OPQ32 report: OPQ32 technical manual supplement*. Thames Ditton, England: SHL Group.

Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ32 technical manual.* Thames Ditton, England: SHL Group.

Bartram, D., & Burke, E. (in press). Industrial/organizational case studies. In J. A. Wollack & J. F. Fremer (Eds.), *Handbook of test security*. New York, NY: Routledge.

Bartram, D., Robertson, I., & Callinan, M. (2002). A framework for examining organizational effectiveness. In I. Robertson, M. Callinan, & D. Bartram (Eds.), *Organizational effectiveness: The role of psychology* (pp. 1–10). Chichester, England: Wiley. doi:10.1002/9780470696736.ch

Bass, B. M. (1997). Does the transactional–transformational leadership paradigm transcend organizational and national boundaries? *American Psychologist, 52*, 130–139. doi:10.1037/0003-066X.52.2.130

Bass, B. M., Burger, P. C., Doktor, R., & Barrett, G. V. (1979). *Assessment of managers: An international comparison.* New York, NY: Free Press.

Benet-Martinez, V. (2007). Cross-cultural personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 170–189). New York, NY: Guilford Press.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107,* 238–246. doi:10.1037/0033-2909.107.2.238

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*, 223–235. doi:10.1111/j.1468-2389.2006.00345.x

Berry, J. W. (1980). Introduction to methodology. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology: Vol. 2. Methodology* (pp. 1–28). Boston, MA: Allyn & Bacon.

Burke, E. (2006). *Better practice for online assessment*. Thames Ditton, England: SHL Group.

Burke, E. (2008a, July). *Applying data forensics to defend the validity of online employment tests.* Paper presented at the Conference of the International Test Commission, Liverpool, England.

Burke, E. (2008b). Coaching with the OPQ. In J. Passmore (Ed.), *Psychometrics in coaching* (pp. 87–114). London, England: Kogan Page.

Burke, E. (2008c, March). *Dealing with the security of online employment testing.* Case study presentation at the Association of Test Publishers Test Security Summit, Dallas, Texas.

Burke, E. (2008d, April). *Preserving the integrity of online testing.* In N. T. Tippins (Chair), Internet testing: Current issues, research solutions, guidelines, and concerns. Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, California.

Burke, E. (2009a, July). From norms to benchmarks and the development of indices of human capital rather than just reference points for test scores. In E. Burke (Chair), *Reconsidering norms: Their construction and their value in enabling valid decisions*. Symposium held at the European Congress of Psychology, Oslo, Norway.

Burke, E. (2009b). Preserving the integrity of online testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 35–38. doi:10.1111/j.1754-9434.2008.01104.x

Burke, E. (2009c, February). *Test development in a global economy*. Workshop delivered at the Innovations in Testing Conference of the Association of Test Publishers, Palm Springs, California.

Burke, E., Bartram, D., & Philpott, D. (2009). *OPQ32i international norm OPQ32: Technical manual supplement.* Thames Ditton, England: SHL Group.

Burke, E., Bartram, D., Wright, D., Rebello, C., & Johannesson, H. (2008, June). *Test adaptation: Lessons learned from a scale multi-language programme.* Workshop delivered at the International Test Commission Conference, Liverpool, England.

Burke, E., & Bateson, J. (2009, April). Technology assisted test construction, delivery and validation. In J. Weiner (Chair) *Technology-based assessment in the 21st century: Advances and Trends.* Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Burke, E., Mahoney-Phillips, J., Bowler, W., & Downey, K. (2011). Going online with assessment: Putting the

science of assessment to the test of client need and 21st century technologies. In N. Tippins & S. Adler (Eds.), *Using technology to enhance the assessment of talent* (pp. 355–379). San Francisco, CA: Jossey-Bass. doi:10.1002/9781118256022.ch14

Burke, E., van Someren, G., & Tatham, N. (2006). *Verify range of ability tests: Technical manual.* Thames Ditton, England: SHL Group.

Burke, E., & Vaughan, C. (2010). *Assessment in a global context: Best practice in localizing and adapting assessments in high stakes scenarios.* Thames Ditton, England: SHL Group.

Burke, E., & Vaughan, C. (2011, April). The generalizability of a construct driven approach to SJTs. In F. Lievens and T. Rockstuhl (Chairs), *Innovation in SJT technology: Item development, fidelity, and constructs assessed.* Symposium presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Burke, E., Vaughan, C., & Ablitt, H. (2010). *Dependability and safety instrument V1.1 technical manual.* Thames Ditton, England: SHL Group.

Burke, E., Vaughan, C., & Ablitt, H. (2011, January). *How to build a global situational judgment test that delivers real value to the recruiter and the candidate.* Workshop delivered at the annual conference of the British Psychological Society's Division of Occupational Psychology, Stratford-upon-Avon, England.

Burke, E., Vaughan, C., & Fix, C. (2009a, July). *Localisation and adaptation of employment tests.* Symposium delivered at the European Congress of Psychology, Oslo, Norway.

Burke, E., Vaughan, C., & Fix, C. (2009b, July). *Situational judgment tests.* Workshop delivered at the European Congress of Psychology, Oslo, Norway.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming.* Mahwah, NJ: Erlbaum.

Davey, T., & Nering, M. (2002). Controlling item exposure & maintaining item security. In C. G. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165–191). Mahwah, NJ: Erlbaum.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73*, 1246–1256. doi:10.1037/0022-3514.73.6.1246

Drasgow, F., Nye, C. D., Guo, J., & Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 46–48. doi:10.1111/j.1754-9434.2008.01106.x

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327–358. doi:10.1037/h0061470

Flynn, J. R. (2007). *What is intelligence?* Cambridge, England: Cambridge University Press.

Geisinger, K. F. (2003). Testing and assessment in cross-cultural psychology. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 12. Assessment psychology* (pp. 95–117). Hoboken, NJ: Wiley.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hofstede, G. (2001). *Culture's consequences: Comparing values, institutions, and organizations across nations.* Thousand Oaks, CA: Sage.

Hofstede, G., & Hofstede, G. J. (2005). *Cultures and organizations: Software of the mind.* New York, NY: McGraw-Hill.

Hogan, J., Hogan, R., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology, 69*, 167–173. doi:10.1037/0021-9010.69.1.167

House, R. J. (2004). Illustrative examples of GLOBE findings. In R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, & V. Gupta (Eds.), *Culture, leadership, and organizations: The GLOBE study of 62 societies* (pp. 3–8). Thousand Oaks, CA: Sage.

House, R. J., & Javidan, M. (2004). Overview of GLOBE. In R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, & V. Gupta (Eds.). *Culture, leadership, and organizations: The GLOBE study of 62 societies* (pp. 9–28). Thousand Oaks, CA: Sage.

House, R. J., Wright, N. S., & Aditya, R. N. (1997). Cross-cultural research on organizational leadership: A critical analysis and a proposed theory. In P. C. Earley & M. Erez (Eds.), *New perspectives in industrial/organizational psychology* (pp. 535–625). San Francisco, CA: New Lexington Press.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–879. doi:10.1037/0021-9010.85.6.869

Impara, J. C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Mahwah, NJ: Erlbaum.

International Test Commission. (2005). *International guidelines on computer-based and Internet delivered testing.* Retrieved from http://www.intest.org/Downloads/ITC%20Guidelines%20on%20Computer%20-%20version%202005%20approved.pdf

International Test Commission. (2010). *International Test Commission guidelines for translating and adapting*

tests: Version 2010. Retrieved from http://www.intest.com.org/upload/sitefiles/40.pdf

Jahoda, G. (1995). In pursuit of the emic–etic distinction: Can we ever capture it? In N. R. Goldberger & J. B. Veroff (Eds.), *The culture and psychology reader* (pp. 128–138). New York, NY: New York University Press.

Judge, T. A., & Cable, D. M. (1997). Applicant personality, organizational culture, and organizational attraction. *Personnel Psychology, 50*, 359–394. doi:10.1111/j.1744-6570.1997.tb00912.x

Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology, 87*, 797–807. doi:10.1037/0021-9010.87.4.797

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology, 84*, 817–824. doi:10.1348/0963179 10X522672

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91*, 247–258. doi:10.1037/0021-9010.91.2.247

Lonner, W. J. (1979). Issues in cross-cultural psychology. In A. J. Marsella, R. G. Tharp, & T. J. Ciborowski (Eds.), *Perspectives on cross-cultural psychology* (pp. 17–45). New York, NY: Academic Press.

Maynes, D. (2009). *Combining statistical evidence for increased power in detecting cheating*. Retrieved from http://caveon.com/articles/Combining_Statistical_Evidence_for_Increased_Power_in_Detecting_Cheating_2009_Apr_04.pdf

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91. doi:10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740. doi:10.1037/0021-9010.86.4.730

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741

Messick, S. (1998). Alternative modes of assessment, uniform standards of validity. In M. D. Hakel (Ed.), *Multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 59–74). Mahwah, NJ: Erlbaum.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low fidelity simulation. *Journal of Applied Psychology, 75*, 640–647.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027. doi:10.1111/j.1744-6570.2007.00099.x

Pike, K. L. (1967). *Language in relation to a unified theory of the structure of human behaviour*. The Hague, the Netherlands: Mouton.

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology, 74*, 441–472. doi:10.1348/096317901167479

Roe, R. (2009, July). Developing a model of the factors influencing test scores and norms. In E. Burke (Chair), *Reconsidering norms: Their construction and their value in enabling valid decisions*. Symposium held at the European Congress of Psychology, Oslo, Norway.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Pierre, J. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology, 88*, 1068–1081. doi:10.1037/0021-9010.88.6.1068

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274. doi:10.1037/0033-2909.124.2.262

Schmitt, D. P., Allik, J., McRae, R. R., & Benet-Martinez, V. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*, 173–212. doi:10.1177/0022022106297299

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135–155). Mahwah, NJ: Erlbaum.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*, 437–453. doi:10.1111/j.1744-6570.1987.tb00609.x

Schwartz, S. H. (1994). Are there universal aspects in the content and structure of values? *Journal of Social Issues, 50*, 19–45. doi:10.1111/j.1540-4560.1994.tb01196.x

Schwartz, S. H., & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. *Journal*

of Personality and Social Psychology, 58, 878–891. doi:10.1037/0022-3514.58.5.878

Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality, 38*, 230–255. doi:10.1016/S0092-6566(03)00069-2

Silzer, R., & Church, A. H. (2009). The pearls and perils of identifying potential. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 377–412. doi:10.1111/j.1754-9434.2009.01163.x

Stokes, G., Mumford, M., & Owens, W. (1994). *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction.* Palo Alto, CA: Consulting Psychologists Press.

Tate, L., & Hughes, D. (2007, January). *To cheat or not to cheat: Candidates' perceptions and experiences of unsupervised computer-based testing.* Paper presented at the conference of the British Psychological Society's Division of Occupational Psychology, Bristol, England.

Taylor, P. J., Pajo, K., Cheung, G. W., & Stringfield, P. (2004). Dimensionality and validity of a structured reference check procedure. *Personnel Psychology, 57*, 745–772. doi:10.1111/j.1744-6570.2004.00006.x

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500–517. doi:10.1037/0021-9010.88.3.500

Tett, R. P., Fitzke, J. R., Wadlington, P. L., Davies, S. A., Anderson, M. G., & Foster, J. (2009). The use of personality test norms in work settings: Effects of sample size and relevance. *Journal of Occupational and Organizational Psychology, 82*, 639–659. doi:10.1348/096317908X336159

Tippins, N. T. (2008, April). *Internet testing: Current issues, research solutions, guidelines, and concerns.* Symposium presented at the annual conference of the Society for Industrial and Organizational Psychology, San Francisco, California.

Tippins, N. T., Beatty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology, 59*, 189–225. doi:10.1111/j.1744-6570.2006.00909.x

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Warr, P., & Pearce, A. (2004). Preferences for careers and organizational cultures as a function of logically related personality traits. *Applied Psychology: International Review, 53,* 423–435.

Weekley, J. A., & Ployhart, R. E. (2006). *Situational judgment tests: Theory, measurement, and application.* Mahwah, NJ: Erlbaum.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182). Mahwah, NJ: Erlbaum.

Weick, K. E. (2001). *Making sense of the organization.* Oxford, England: Blackwell.

Wright, D. (2009, July). Applying meta-regression to norms and their interpretation. In E. Burke (Chair), *Reconsidering norms: Their construction and their value in enabling valid decisions.* Symposium held at the European Congress of Psychology, Oslo, Norway.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

CHAPTER 33

PERFORMANCE APPRAISAL

Kevin R. Murphy and Paige J. Deckert

In work organizations of all sorts (e.g., private sector, public sector, military), it is useful and often necessary to measure the performance and effectiveness of individuals and teams. This measurement is usually done via performance appraisals that are based largely on the judgments of supervisors, peers, customers, or some other evaluators. Other methods might be used to measure job performance (e.g., objective productivity counts), but measures based on judgments and subjective evaluations of performance are much more common. Supervisory performance appraisals are virtually universal in public-sector organizations (except in specific circumstances in which union contracts or regulations call for other approaches), and they are used in the great majority of moderately large and large organizations. They are less likely to be formalized in small businesses, but even in small organizations, it is common to provide employees with annual feedback about their performance and effectiveness.

Performance appraisals are often an important factor in decision about pay, promotion, and developmental opportunities (Landy & Conte, 2007). They are an important (but not always welcome) source of feedback (Cleveland, Murphy, & Lim, 2007; Kluger & DeNisi, 1996; Leung, Su, & Morris, 2001). Performance appraisals are often at the heart of equal employment litigation, particularly in cases in which a plaintiff claims to have been evaluated unfairly (Barrett & Kernan, 1987; Cascio & Bernardin, 1981). They are widely used as criteria for validating personnel tests (Landy & Farr, 1980).

Performance appraisal represents a method of measurement that depends on informed evaluative judgments (Milkovich & Wigdor, 1991). That is, performance appraisal systems are usually designed around the assumption that the judges who are called on to evaluate performance (the term *raters* is used here) have access to information about the performance of the individuals they evaluate (*ratees*) and have an understanding of the appropriate standards that should be used in determining whether performance is adequate, exemplary, or inadequate. Substantial bodies of research have examined the extent to which each of these assumptions is met (Milkovich & Wigdor, 1991), and it has been shown that appraisal systems that depend on raters who are uninformed or whose judgments cannot be trusted or calibrated are unlikely to provide good measures of performance.

There is a long history of dissatisfaction and concern with performance appraisal (Austin & Villanova, 1992; McGregor, 1957; Patz, 1975). Neither raters nor ratees are likely to report a great deal of support for or trust in performance appraisals (Boswell & Boudreau, 2000; Murphy & Cleveland, 1995; Tziner & Murphy, 1999; Tziner, Murphy, & Cleveland, 2001; Tziner, Murphy, Cleveland, Beaudin, & Marchand, 1998). Ratees are sometimes unwilling to seek or accept feedback about their performance unless they are confident that the feedback will be positive (Ashford, Blatt, & VandeWalle, 2003).

Some authors have gone as far as suggesting that performance appraisal should be abolished (Coen & Jenkins, 2000). This pessimism may not be warranted. The costs of doing performance appraisal can exceed the benefits, especially when appraisals are done poorly (Murphy & Cleveland, 1995), but if

DOI: 10.1037/14047-033
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

appraisals are done with care and used sensibly, they can be use useful to ratees and organizations alike.

This chapter is organized around three key questions in performance appraisal: who, how, and why. The chapter examines who conducts and who receives performance appraisals in work settings (and who does not), how these appraisals are done and evaluated, and how they are used and why they are necessary. Four key themes run through the research on performance appraisal and are discussed in this chapter:

1. Performance appraisal requires informed judgment (i.e., an assessment based on a knowledge of the job and a knowledge of ratee's behavior). Elaborate training programs, appraisal forms, rating processes, and cross-checks are aids to judgment but are not substitutes for it.
2. The way appraisals are used (both formally and informally) in organizations has a substantial effect on both the process and outcomes of appraisal.
3. Performance appraisals are more than simply an exercise in performance measurement. The motivation and goals of raters and ratees and the context in which appraisals are conducted are crucial determinants of performance ratings and the effectiveness of performance appraisal systems.
4. Performance appraisal can unfortunately be thought of as a well-developed, carefully instrumented system for making people unhappy. Because of well-established differences in self-evaluations and evaluations received from others, most people are likely to be dissatisfied with the ratings they receive.

Before discussing performance appraisal in detail, it is useful to understand how performance appraisal fits in the systems that organizations typically develop and use to manage employee performance. Figure 33.1 illustrates a typical performance management system.

Most organizations use a performance management system that starts with a process of setting goals and objectives, which might be negotiated between supervisors and subordinates or might be imposed by supervisors. Either way, the first step in performance management is a specification of what employees are expected to do and accomplish. This step leads to a process of collecting and integrating information to evaluate performance. Performance evaluations often include a mix of objective information about goal accomplishment and evaluations provided by supervisors. The results of these evaluations are usually fed back to subordinates in some way, and this performance feedback is often folded into a set of action plans for improving or maintaining performance levels.

The focus of this chapter is on the second step in Figure 33.1, evaluation, though the third step, feedback, is also touched on. Assessments of the validity and impact of performance appraisals sometimes

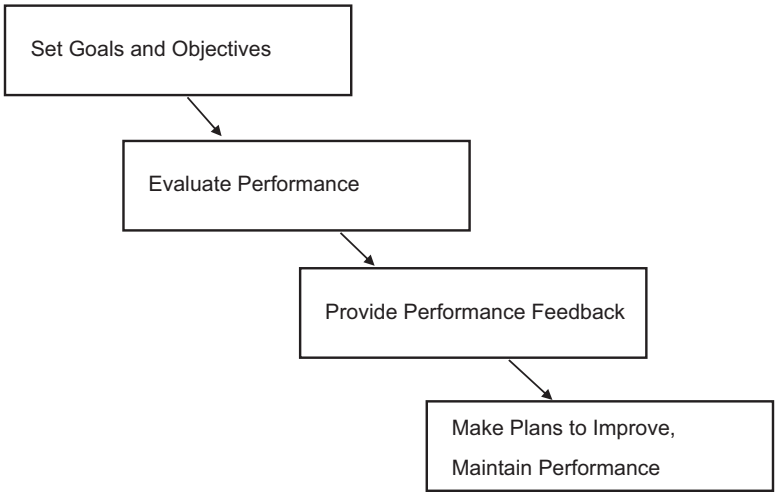

FIGURE 33.1. Performance management systems.

draw on assessments of the success or failure of action plans that are formed on the basis of performance evaluations, and the job relatedness of performance evaluations often depends heavily on the process by which goals and objectives are formed and communicated.

WHO GIVES AND RECEIVES APPRAISALS?

Several sources can be used in evaluating performance, including supervisors, peers, subordinates, clients and customers, and even the ratee. The most common arrangement involves obtaining performance judgments from the ratee's direct supervisor or manager (Landy & Farr, 1980); one of the defining characteristics of a supervisory relationship is that the supervisor has the right and the responsibility to evaluate his or her subordinates (Dornbusch & Scott, 1975). In the past 10 years, the emphasis on the use of multiple sources for evaluation has been increasing, particularly on 360-degree appraisal systems (Atwater, Waldman, & Brett, 2002; Bracken, Timmreck, & Church, 2000). Most 360-degree systems are used to provide feedback to ratees and are based on the assumption that supervisors, peers, subordinates, and other sources all have potentially useful (and potentially unique) things to say about the strengths and weaknesses of ratees. A 360-degree system obtains input from a number of individuals at different levels of an organization and typically provides feedback to individual ratees about how their performance is viewed by supervisors, peers, subordinates, and others.

One particularly difficult problem with multisource systems for performance evaluation is that raters who occupy different roles in organizations (e.g., supervisors vs. peers) often disagree in their evaluations of performance (Conway & Huffcutt, 1997; Murphy, Cleveland, & Mohler, 2001). For example, supervisors see different behaviors than are seen by peers and are likely to evaluate them from a different frame of reference. Similarly, peers see different behaviors and evaluate them differently than subordinates (Murphy & Cleveland, 1995). Even if one focuses solely on raters who are at a comparable level in the organization (e.g., ratings from multiple supervisors), it is likely that different individuals will, as a result of their position in the organization, work assignments, differences in perspective, and so forth, see different samples of behavior and evaluate those behaviors differently. Unfortunately, as Murphy et al. (2001) noted, inconsistency and disagreement often lead recipients to dismiss the feedback they receive, especially if it is negative.

Two trends are well established in research examining the circumstances under which raters are likely to agree or disagree in their evaluations. First, self-ratings of performance are usually higher, or more lenient, than ratings obtained from supervisors and peers (Farh & Werbel, 1986; Harris & Schaubroeck, 1988; Thornton, 1980). This discrepancy is a constant source of trouble in organizations because many ratees believe that their supervisors and peers underestimate their level of performance. Second, raters who are at similar levels in organizations, and should therefore have similar perspectives, nevertheless often disagree in their evaluations. Conway and Huffcutt's (1997) meta-analysis showed that subordinates had the lowest level of interrater reliability. On average, subordinate ratings of job performance show correlations in the low .30s; average interrater correlations are slightly higher for peers (.37). Supervisors show slightly higher levels of agreement (.50s; see also Viswesvaran, Ones, & Schmidt, 1996), but in general, similarly situated raters tend to provide evaluations that are at best moderately consistent.

The research literature on performance appraisal often presents a simplistic description of the sources of information most likely to be used in performance appraisal—for example, supervisor versus self versus 360 degree. In most moderately large organizations (performance appraisal practices in small businesses are both less formal and less well understood than in larger organizations), mixed systems prevail. It is common for subordinates to have input into performance appraisals, particularly in the process of setting and evaluating progress toward achieving key goals. It is also common for supervisory judgments to be reviewed, and sometimes even revised, by upper management or through some sort of peer comparison system (e.g., the rating distributions of different managers might

be compared). Thus, although people typically think of performance appraisals as supervisory judgments, they are judgments that incorporate information from multiple sources, and they are often subject to organizational checks and balances.

Although upward evaluation (e.g., subordinates evaluating their supervisors) is often a component of performance feedback systems, there is little doubt that downward evaluations are taken more seriously than upward evaluations (Murphy & Cleveland, 1995; Murphy et al., 2001). Organizations are, after all, hierarchical, and the decisions and evaluations of individuals at higher levels almost always carry more weight than those obtained from those at lower levels of the organization.

Not only do organizational levels affect the weight given to performance evaluations, they also influence the nature of the appraisal process. On the whole, performance appraisals conducted lower in the organization tend to be simpler and more uniform, often using structured performance appraisal forms to obtain judgments from a single supervisor. Appraisals of supervisors and managers are often more complex and less structured, often incorporating significant input from the ratee as well as a mix of judgments and outcome measures.

One paradox is that higher level executives, whose actions and effectiveness could arguably have substantial impact on the organization, are least likely to receive formal performance appraisals. At lower levels of the organization, appraisal is often an annual occurrence, with careful records and cross-checking of supervisors' judgments (e.g., many appraisals require sign off from higher level managers). High-level executives may not receive any formal appraisals, and the appraisals they do receive may be quite informal and unstructured. Given the potential importance of effective performance on the part of higher level executives, organizations would be well served to devote time, structure, and attention to evaluating executive performance.

HOW ARE APPRAISALS CONDUCTED?

The way in which performance appraisals are conducted in organizations often reflects the purposes of appraisal and the way information from appraisals will be used to make decisions about individuals or organizational programs and practices. Most organizations conduct annual evaluations of each employee, tied to administrative decision cycles (e.g., pay determination). Appraisals that are done for feedback purposes might be more frequent and less regular in their schedule. Appraisals that are used as criteria for evaluating training programs or personnel selection systems may be done separately from those that are done for the purpose of making pay and promotion decisions (Meyer, Kay, & French, 1965).

The prototypic performance appraisal in organizations is

- conducted by an individual supervisor, who is asked to make judgments about overall performance levels and about separate facets or dimensions of performance (e.g., problem solving, oral communication);
- based on personal observations, often supplemented by input from the ratee, descriptions of goals and accomplishments, and input from other members of the organization; and
- reviewed by superiors of that supervisor (which might be nothing more than a nominal check off) and used as a input for making a range of decisions about the rate (e.g., pay increases, promotions).

Performance judgments might come in many different forms, the most common of which would include ratings, rankings, or forced distributions. Rating systems ask supervisors to compare each ratee with a standard (e.g., performance meets expectations, performance is above average). Ranking systems ask supervisors to compare ratees with one another (e.g., "of the four employees whom I supervise, Fred is the second best performer"). Forced distribution systems ask supervisors to sort ratees into ordered categories, usually with some sort of quota (e.g., sort employees into A [top 20%], B [middle 60%], and C [bottom 20%] categories).

Forced distribution systems are favored by some organizations (Welch, 2001), but these systems are most likely to be controversial, especially when rewards and sanctions are closely tied to the category to which each ratee is assigned. For example, Welch (2001) advocated a "rank-and-yank" system,

in which the weakest performers (often the lowest 10%) are identified and dismissed (Scullen, Bergey, & Aimon-Smith, 2005). This system is often viewed as arbitrary, unfair, and subjective, especially if the weaker employees are performing adequately, just not as well as some of their peers (Roch, Sturnburgh, & Caputo, 2007). An important limitation of the rank-and-yank approach is that after a certain period of time, the poorest performer in the company will be equal to the best applicant and better than most applicants. Replacing the poorest performers with new hires will therefore result in a loss in performance, not a gain. A simulation of the effects of rank and yank found that the first firm experienced a decrease in performance at 9 years when assuming no voluntary turnover (the ideal situation) and as early as 4.5 years with 20% voluntary turnover (Scullen et al., 2005).

Ranking systems are less common, especially when the number of employees to be compared is more than a handful. Ranking systems are common in the military, but in the private and public sectors, rating systems are by far the most common (Landy & Farr, 1980, 1983).

As Figure 33.1 suggests, the starting point for most performance appraisals involves supervisors and subordinates working together to set goals and objectives. This focus on concrete, observable goals and objectives, which originated as a component of the management by objectives process, defines what employees are expected to accomplish over specific time periods and also defines the relevant dimensions and criteria for evaluating performance at the end of that period. By specifying and agreeing on goals and objectives ahead of time, both supervisors and subordinates can remove a good deal of the uncertainty and subjectivity that is sometimes characteristic of performance appraisal. Appraisals are not necessarily limited to counts of goals met and objectives achieved, but modern performance appraisal systems will typically start with this goal-setting process and use the accomplishment of goals as a centerpiece of performance appraisal.

What to Rate

A number of taxonomies have been used to describe the major dimensions of the job performance domain (e.g., Borman & Brush, 1993; Campbell, 1990; Hunt, 1996; Murphy, 1989). Borman, Bryant, and Dorio (2010) summarized the major dimensions of performance across a wide range of jobs in terms of

- communication and interaction,
- productivity and proficiency,
- problem solving,
- information processing,
- organizing and planning,
- leadership and supervision,
- counterproductive work behaviors, and
- useful personal qualities (e.g., conscientiousness, initiative).

Most generally, performance ratings reflect two broad domains of behavior, task performance and organizational citizenship behaviors (OCBs; Borman & Motowidlo, 1993; Podsakoff, MacKenzie, Paine, & Bachrach, 2000). *OCB* is defined as "individual behavior that is discretionary, not directly or explicitly recognized by a formal rewards system, and that in aggregate promotes the effective functioning of the organization" (Organ, 1988, p. 4). The original conceptualization of OCB encompassed five dimensions: conscientiousness, sportsmanship, altruism, courtesy, and civic virtue (Organ, 1988). Other conceptualizations of OCB have distinguished between behaviors directed toward the organization and behaviors directed toward individuals (e.g., coworkers; Williams & Anderson, 1991), although some meta-analyses question the utility of this distinction (e.g., LePine, Erez, & Johnson, 2002). In the study of OCB, researchers have been anything but consistent. In the 133 studies surveyed by LePine et al. (2002), more than 40 measures of OCB or OCB-related behaviors were reported. Although it may be difficult to develop a comprehensive measure of OCBs, the Organ (1988) conceptualization of this construct still suggests some useful dimensions that are likely to run through the evaluation of OCBs in most jobs, particularly conscientiousness, sportsmanship, and courtesy, which are likely to be relevant across a wide range of social situations encountered in the workplace.

How to Rate

A substantial research literature has dealt with the advantages and disadvantages of various rating scale

formats. Landy and Farr (1980) noted that the practical effects of using different types of rating scales are often quite small, and they even called for a moratorium on rating scale format studies (see also Bernardin, Alvares, & Cranny, 1976). Nevertheless, there are some good reasons to pay attention to rating scales. At their best, well-designed rating scales may help to elicit consistent and well-informed judgments, whereas poorly defined scales may simply add to the confusion and disappointment that often surrounds performance appraisal (Borman et al., 2010; Landy & Farr, 1980, 1983).

Graphic rating scales. The simplest scale format, the graphic rating scale asks the rater to record his or her judgment about some specific aspect of the ratee's performance on a scale that can be used to obtain numeric values that correspond with the rater's evaluation of the ratee. Several examples of a graphic scale that might be used to record ratings of a performance dimension such as oral communication are presented in Figure 33.2.

This type of scale format provides little structure for the rater in recording his or her judgment. Graphic scales can range from those such as the first scale, which contain no definitions of what is meant by poor, good, or average levels of performance, to those that define each level in terms of some label (second scale), or even in terms of a brief description of what is meant by each level of performance (third scale).

The principal advantage of this scale type is simplicity. The disadvantage of this format, which led to efforts to develop alternative formats, is the lack of clarity and definition. First, the scales do not do much to define what the dimension oral communication means. Different supervisors might include very different behaviors, interactions, and so forth under this general heading. Second, the scales do not do much to define what is meant by *poor, average,* and so forth. Supervisors might apply a variety of different standards when evaluating the same behaviors. Many behavioral-based scale formats were developed in an effort to solve these problems, in particular behaviorally anchored rating scales and behavioral observation scales.

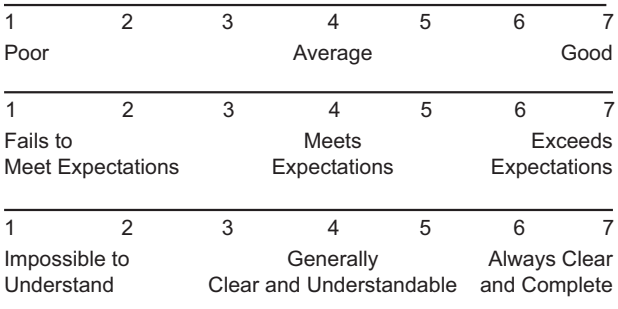

FIGURE 33.2. Graphic rating scales for evaluating oral communication.

Behaviorally anchored rating scales. The development and use of behaviorally anchored rating scales (BARS) accounted for much of the research on performance appraisal scales in the 1960s through the 1980s (e.g., Bernardin, 1977; Bernardin & Smith, 1981; Borman, 1986; Jacobs, Kafry, & Zedeck, 1980; Landy & Barnes, 1979; Smith & Kendall, 1963). These scales use behavioral examples of different levels of performance to define both the dimension being rated and the performance levels on the scale in clear, behavioral terms. The process of scale development can be long and complex, but it will usually result in scales that are clearly defined and well accepted by both raters and ratees.

An example of a BARS scale similar to those used by Murphy and Constans (1987) in one of their studies of teacher rating is presented in Figure 33.3. The behavioral examples are designed to do two things: (a) illustrate clearly what *oral communication* means and (b) illustrate clearly and concretely what good, average, and poor performance might look like. The theory of BARS suggests that this type of scale can be an effective tool for helping ensure that raters adopt a common frame of reference.

Much of the rating format research of the 1970s seemed to reflect the assumption that BARS were more objective than graphic scales and that defining performance in behavioral terms would result in more accurate ratings. This assumption was not supported in subsequent research (Landy & Farr, 1980; Murphy & Constans, 1987; Murphy, Martin, & Garcia, 1982), which has led many researchers and practitioners to question the utility of BARS. This question is especially relevant because the process of developing BARS can be time consuming and

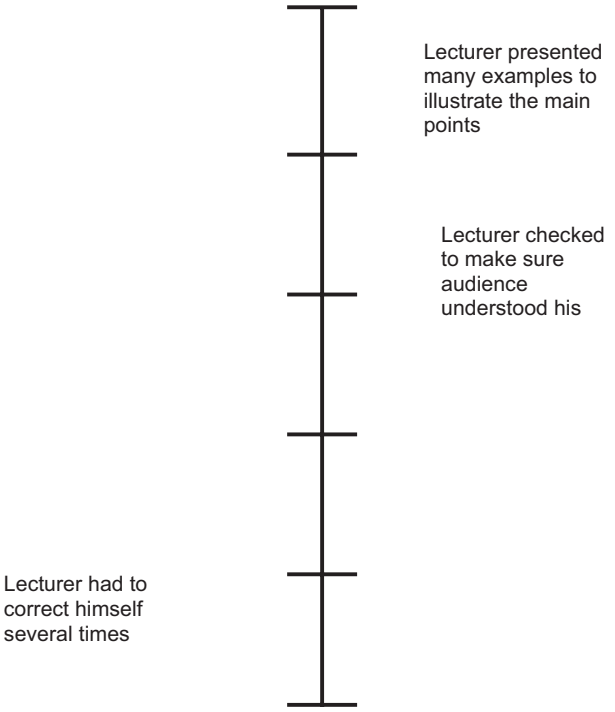
Lecturer presented many examples to illustrate the main points
Lecturer checked to make sure audience understood his
Lecturer had to correct himself several times

FIGURE 33.3. Behaviorally anchored scale for rating oral communication.

expensive (Smith & Kendall, 1963). However, BARS appear to have one advantage that was not fully anticipated by early BARS researchers—that is, they are accepted by users. The reason for this acceptance is that most BARS development procedures incorporate feedback from large numbers of raters (and, sometimes, ratees) in the process of constructing scales. As a result, many of the raters and ratees are likely to feel that they have some personal investment in the scales. Even those raters and ratees who do not participate in scale development may view the scales favorably because of the heavy reliance on their colleagues' feedback into numeric scores (Bernardin & Beatty, 1984).

Behavior observation scales. A final variation on the use of behavioral examples in evaluating performance is the behavior observation scale (BOS; Latham, Fay, & Saari, 1979). The BOS asks the rater to describe how frequently each behavior occurred over the time period covered by appraisal. Proponents of BOSs have suggested that this method removes much of the subjectivity that is usually present in evaluative judgments. Unfortunately, research into the cognitive processes involved in responding to BOSs (Murphy & Constans, 1987; Murphy et al., 1982) has suggested that the process of judging behavior frequency is every bit as subjective as the process of forming evaluative judgments. Behavior frequency ratings may in fact be more subjective than trait ratings or overall judgments; overall evaluations of the ratee's performance appear to serve as a critical cue for estimating behavior frequencies. Thus, the use of BOSs probably does not allow one to avoid the subjectivity of overall impressions or judgments.

This chapter describes potential advantages in using graphic scales and BARS. We are not as enthusiastic about BOSs. The behavioral orientation of these scales appears, on the surface, to be a decided advantage, but there are several reasons to believe that raters do not respond to these scales in terms of behaviors. Rather, they use their overall, subjective evaluations to guide their behavior ratings. This type of scale might actually disguise the inherent subjectivity of evaluative judgment by phrasing judgments in an apparently objective behavioral language. The rather negative evaluation of this type of scale may reflect our biases as much as it reflects the shortcomings of BOSs. One of the authors (Kevin R. Murphy) has been involved in much of the research questioning BOSs, and it is possible that other researchers do not share this evaluation of BOSs.

Performance distribution assessment. Performance distribution assessment (PDA) represents a more sophisticated version of the basic approach exemplified by BOSs (Kane, 1986). In PDA, raters must indicate the frequency of different outcomes (e.g., behaviors, results) that indicate specific levels of performance on a given dimension. For example, the scale might describe the most effective outcome and the least effective outcome that could reasonably be expected in a particular job function as well as several intermediate outcomes. The rater is asked to estimate the frequency of each outcome level for each ratee. One of the potential advantages of this format is that it allows one to consider the distribution or the variability of performance as well as the average level of performance in forming an evaluation. PDA involves some fairly complex scoring rules (a concise description of PDA

is presented in Bernardin & Beatty, 1984; software now exists for PDA scoring) and results in measures of the relative effectiveness of performance, the consistency of performance, and the frequency with which especially positive or negative outcomes are observed.

The evaluation of PDA is similar to the evaluation of BOSs presented earlier. Both depend on the rater's ability to accurately indicate the frequency of specific behaviors or outcomes. Cognitive research has suggested that raters are simply incapable of performing this task in an objective way. It is very likely that raters infer the frequency of different behaviors or outcomes from their global evaluations of individuals and that when one asks for data on the frequency of effective or ineffective behaviors, what one actually gets is a restatement of the rater's overall evaluation. Thus, we do not believe that assessments obtained using PDA or BOSs will be more specific, objective, or behavior based than assessments obtained with much simpler scales.

Employee comparison methods. There is a useful distinction between *rating* and *ranking* (i.e., employee comparison). Rating involves comparing a person with a standard. This standard might be undefined or subjective (e.g., a scale on which the anchors of *good, average,* and *poor* are undefined), or it might be defined in exact behavioral terms. Ranking involves comparing a person with another person. Evidence has shown that the psychological processes involved in rating versus ranking may be different (Murphy & Constans, 1987). Even if this is true, however, ratings and rankings often lead to similar conclusions about the performance of a group of rates (Murphy & Cleveland, 1995).

To illustrate ranking procedures, consider the example of a supervisor who evaluates eight subordinates. One possibility is to simply rank order these eight individuals from the best performer to the worst. With a small number of ratees, this task should not be difficult. However, as the supervisor's span of control increases, ranking of all subordinates can become tedious and sometimes arbitrary. Although it might be easy to pick the best and the worst performers out of a group of 30 workers, it can be very difficult to distinguish the 15th best from the 16th, 17th, or 18th. The forced-distribution ranking procedure provides a partial solution to this problem.

A forced-distribution scale requires supervisors to sort subordinates into ordered categories, such as top performers (e.g., top 20%), average performers (e.g., middle 60%), and poorer performers (e.g., bottom 20%). The principal distinction between a forced-distribution scale and a scale that requires one to rank all subordinates is that in a full ranking the number of categories is equal to the number of people being evaluated. In a forced-distribution scale, the number of categories is less than the number of people. The choice between these two methods depends in part on the specificity of the information required. If there are different outcomes for each individual (e.g., the sixth-best performer will get a larger raise than the seventh-best performer), full ranking is worthwhile. Otherwise, a forced-distribution scale might be easier to use.

If the rater or the organization requires precise information about the rank ordering of employees and the size of the differences in performance among employees, one more procedure can be considered. The pair-comparison method allows the scaling of subordinates with some precision on a ratio-level scale of overall performance. As the name implies, this method required raters to compare each pair of ratees, each time indicating which ratee is the better of the two in performing his or her job. If the number of comparisons is sufficiently large, scaling procedures can be applied that transform these pairwise comparisons into a ratio scale that establishes both the ranking and the extent to which subordinates differ in their performance. The principal drawback of this method is that the number of comparisons expands geometrically as the number of subordinates increases. In the earlier example, six comparisons were needed to evaluate four subordinates. If there were 10 subordinates, 45 comparisons would be needed. With 20 subordinates, 190 comparisons are needed. This assumes, however, that every possible comparison is made; there are research designs that allow inferences to be made about comparisons that are (by design) omitted from the data collection process (Morales & Bautista, 2008).

The users of a pair-comparison scale face a real dilemma. The accuracy of the scaling is a direct function of the number of comparisons that are made. Thus, if the number of comparisons is sufficiently small to be easily carried out, the scaling may not be precise. If the number of comparisons is sufficiently large to yield accurate measurement, the pair-comparison procedure may be extremely time consuming. For this reason, pair-comparison procedures seem to have attracted more attention in the basic research literature than in the field.

Objective Measures of Job Performance

Performance measures that require little or no judgment (e.g., production counts) are referred to as *objective,* whereas measures that depend fundamentally on judgment (e.g., supervisory ratings) are referred to as *subjective.* The terms *objective* and *subjective* are best thought of as endpoints of a continuum rather than as a dichotomy (Borman et al., 2010; Landy & Farr, 1983); except in the most trivial cases, judgment is likely to play some role in virtually every performance measure. Many performance appraisals include a mix of relatively objective measures (e.g., measures of goal accomplishment) and relatively subjective ones. Objective and subjective measures of performance typically show at least modest levels of correlation (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Heneman, 1986) and often show correlations comparable to correlations among peer, supervisor, and self-ratings.

Despite the seeming allure of objective job performance measures, few situations arise in which subjective measures can be completely replaced by objective ones. Landy and Farr (1983) noted that many objective measures have surprisingly low levels of reliability and show little consistency across what should be equivalent indices; when examining 40 different measures of absenteeism, they found the correlations across different indices to be almost zero. The main shortcoming of objective measures of job performance, however, is that they almost always have some sort of criterion deficiency (Borman et al., 2010). Objective measures usually capture only a narrow slice of the entire criterion space. For example, qualities such as teamwork and leadership are not easily amenable to objective criteria, and performance measures that ignore these qualities are likely to be deficient, in much the same way that a count of the number of patients a doctor sees in a day would be a deficient measure without taking into account the quality level of care provided. The best use of objective measures is probably in conjunction with judgmental (subjective) measures, in which objective indices are used to assess those aspects of performance that are both countable and important and in which judgments are used to assess those aspects of performance that are not so easy to count.

HOW ARE APPRAISALS USED IN ORGANIZATIONS?

Cleveland, Murphy, and Williams (1989) identified 20 uses of performance appraisals in organizations, ranging from salary administration to promotion, termination, identifying possible goals, and meeting legal requirements. These purposes can be subsumed under four main themes: using appraisal to distinguish among individuals, distinguishing individual strengths from weaknesses, performing system maintenance (e.g., evaluating human resources systems), and providing documentation (Cleveland et al., 1989; Murphy & Cleveland, 1995).

A good deal of evidence has shown that the way appraisals are used in organizations influences ratings. For example, Jawahar and Williams (1997) found that performance ratings used for administrative purposes were an average of 1 standard deviation higher than those used for used for employee development. Murphy and Cleveland (1995) suggested that raters pay careful attention to the relationships between ratings and the rewards or sanctions received by employees and often rate in the way they think will bring about the desired rewards rather than giving ratings that reflect their true evaluations. When performance appraisals are used to distribute rewards, such as pay increases or promotions, raters are likely to be more strategic and political in their ratings than when they are used, for example, to provide feedback (Sims, Gioia, & Longenecker, 1987).

The two most important uses of performance appraisal are to support administrative decisions

(e.g., raises, promotions) and to provide performance feedback. Feedback in performance appraisal relies on three major assumptions: that employees want feedback about their performance, that supervisors can give useful feedback, and that timely and accurate feedback will lead to positive change (Cleveland et al., 2007). Cleveland et al. (2007) noted that none of these assumptions are likely to be warranted because feedback can have a range of negative consequences, including perceptions of unfair treatment, and possible emotional costs for both the employee and the manager. Many factors play into willingness to give or receive feedback, including demographic characteristics, personality and performance levels, and organizational characteristics such as the performance management system and feedback climate (Cleveland et al., 2007).

Despite the difficulties involved in giving and receiving feedback, the evidence of the benefits of performance feedback is clear. Feedback sessions have been shown to increase employee satisfaction with a performance appraisal process (Dorfman, Stephan, & Loveland, 1986). Three key facets about the delivery of feedback are that it is immediate (i.e., it should happen closely after the rating), specific, and supportive (Kluger & DeNisi, 1996; Thornton & Rupp, 2006). Additionally, Murphy and Cleveland (1995) found that frequent appraisals, agreement concerning job duties, and congruency concerning the standards of good and poor performance all lead to high levels of feedback acceptance by ratees.

When giving feedback, it is important to take context and purpose into account (Murphy & Cleveland, 1995). For example, when feedback is used to help ratees set goals for development, providing detail is essential (Atkins, Wood, & Rutgers, 2002). However, it is important to keep in mind the cognitive load on feedback recipients because overloading them with detail could hinder their understanding of the feedback (Atkins et al., 2002). It is best to begin with positive feedback to provide accurate feedback (including negative) without impairing acceptance (Atwater & Brett, 2006). A small amount of negative feedback can lead to improved performance, but a large amount of negative feedback can impair future performance (Smither & Walker, 2004). Finally, individuals are more likely to accept feedback that compares them with a neutral standard than feedback that compares them with peers; the latter can lead to a feeling of competition and in turn hinder acceptance of feedback (Atwater & Brett, 2006).

Traditionally, ratings were hypothesized to not always reflect job performance because of raters' lack of skills, training, knowledge, opportunity to observe performance, and so forth (Murphy, 2008). An alternative explanation, however, proposed by Banks and Murphy (1985) is that raters are not motivated to provide accurate ratings (see also Cleveland & Murphy, 1992; Murphy & Cleveland, 1995; Sims et al., 1987).

The point of departure for many models of performance rating (e.g., DeCotiis & Petit, 1978; DeNisi, 1996; Landy & Farr, 1980; Murphy & Cleveland, 1995; Sims et al., 1987) is the question "What is the rater trying to do when he or she completes a performance rating form?" Harris (1994); Harris, Ispas, and Schmidt (2008); Hollenbeck (2008); and King (2008) suggested that the answer is more complicated than "They are trying their best to measure the performance of their subordinates" (see also Banks & Murphy, 1985; Levy & Williams, 2004). From the beginnings of personnel psychology through the 1970s, researchers and practitioners treated raters as though they were measurement devices. That is, the typical assumption was that raters were trying to measure the performance of their subordinates when completing performance appraisals and that if practitioners could give them better tools (e.g., better scales, training), they would do a better job (Murphy & Cleveland, 1995). A series of papers in the 1970s and 1980s (notably, DeCotiis & Petit, 1978; Landy & Farr, 1980; Sims et al., 1987) challenged this assumption and led to a more nuanced understanding of what raters are doing and why.

Consensus is emerging that raters pursue a variety of goals when completing performance appraisals, and the accurate measurement of ratee performance is unlikely to be their most important goal (Levy & Williams, 2004; Murphy & Cleveland, 1995; Murphy et al., 2004; Sims et al., 1987). The traditional explanation for many of the shortcomings of performance appraisal is that raters are not able to evaluate performance accurately, but it is

likely in many settings that motivation is a more important factor than ability (Banks & Murphy, 1985; Gioia & Longenecker, 1994; Murphy et al., 2004). In seeking to understand the processes underlying performance rating, Bjerke, Cleveland, Morrison, and Wilson (1987) and Sims et al. (1987) did something that is fairly rare in research on performance appraisal—they talked to raters and asked them what they were doing and why. Although self-reports cannot necessarily be taken at face value, it is notable that very few managers report that they do their best to measure subordinate performance accurately when completing performance appraisals. Rather, they report that they provide ratings that they hope will motivate their subordinates, will help to maintain the harmony of the workgroup, or will make them look good to their subordinates (Murphy & Cleveland, 1995).

In the 1990s, Cleveland and Murphy (1992; Murphy & Cleveland, 1995) developed models describing performance appraisal as goal-directed behavior and articulated social, organizational, and environmental factors that could lead raters to pursue different goals when rating their subordinates' performance. In collaboration with Tziner et al. (1998, 2001), Murphy and Cleveland have empirically tested and confirmed many of the predictions of these models (see also Murphy et al., 2004). These models of performance rating processes in organizations have suggested that ratees' performance level does indeed affect performance ratings. Ones, Viswesvaran, and Schmidt (2008) summarized evidence supporting the construct validity of performance ratings. However, performance appraisal cannot be adequately understood as a simple effort to measure job performance. Rather, performance appraisal is a complex event that occurs in environments that often push raters to distort their ratings to accomplish valued goals or to avoid the negative repercussions of giving ratings their subordinates or superiors will find objectionable. The recognition that raters in organizations are not simply passive measurement instruments, trying their best to give accurate measures of their subordinates' performance, is critically important to understanding the potential sources for both systematic variance and error variance in performance ratings.

Tziner and his colleagues (e.g., Tziner et al., 2001) suggested that rater attitudes toward organizations influence performance ratings. For example, raters who perceive a participative organizational climate and a more positive affective commitment to the organization tend to (a) give higher ratings, (b) make smaller distinctions among the subordinates they evaluate, and (c) make stronger distinctions among the strengths and weaknesses of their subordinates. One implication of these findings is that the feedback subordinates receive may partly depend on whether the rater views the organization positively. Supervisors who are strongly invested in the organization and in the concept of participation may be more lenient and less discriminating. However, raters who are disengaged and authoritarian may be harsher and more judgmental.

Studies of the role of organizational factors in perceptions of performance appraisal (e.g., Tziner & Murphy, 2001; Tziner et al., 1998, 2001) have suggested that more proximal attitudes (i.e., perceptions of human resources systems and of the performance appraisal process) have a stronger effect on performance evaluations. In particular, supervisors' beliefs about the way performance evaluations are used in organizations (purpose of rating) and about the way their colleagues conduct performance appraisals (performance appraisal politics) seem particularly important. Supervisors who believe that performance evaluations will be used to make important decisions about their subordinates (e.g., raises, promotions) are likely to inflate ratings. Similarly, supervisors who believe their colleagues manipulate ratings to accomplish political ends (e.g., maintaining harmony in the work group, making the supervisor look good) are likely to inflate ratings. Again, these findings imply that the performance feedback one receives is the result not only of actual employee performance levels, but also of the supervisor's trust or lack of trust in other supervisors and of his or her perceptions of the links between performance ratings and valued outcomes and rewards. Farr and Jacobs (2006) argued that trust drives both employees' perceptions of the performance appraisal system and also the potential outcomes of the system. The relationship to trust in the system holds for both supervisors and

subordinates; supervisors need to be able to put faith in an appraisal system that they perceive as fair and accurate (Mayer, Davis, & Schoorman, 1995), and if the supervisor conducting the appraisal is perceived as trustworthy, subordinates are more likely to trust the fairness of the system (Farr & Jacobs, 2006). Thus, trust in the appraisal system may be an important component of building and maintaining trust between supervisors and subordinates. One potential mechanism for increasing trust is for raters to be honest with ratees about why they are giving specific ratings and to acknowledge what is widely assumed by both raters and ratees—that is, that performance appraisal is more than a simple record of the ratee's behavior and effectiveness. It is a communication to the employee and the organization that is designed to meet a complex set of goals, and performance measurement is only one of these goals (Murphy & Cleveland, 1995).

EVALUATING PERFORMANCE APPRAISALS

The problem of determining whether performance appraisals faithfully reflect the performance levels of the subordinates who are rated has proved to be a difficult one. Historically, indirect criteria have been used. For example, several so-called "rater errors," most notably leniency and halo, have been identified; ratings that are free from these errors are presumed to be better measures than ratings that reflect substantial leniency or halo. In the 1980s, direct measures of rating accuracy were developed, but these measures are only applicable in controlled environments, such as laboratory studies. More current efforts have focused on assessing the reliability and construct validity of ratings.

Leniency error is the tendency for the rater to provide inflated ratings. For example, Bretz, Milkovich, and Read (1992, p. 333) concluded that "the norm in U.S. industry is to rate employees at the top end of the scale." In fact, it is common for 60% to 70% of the workforce to be categorized into the top two levels of performance (Bretz et al., 1992). Bernardin and Orban (1990) found that the degree of trust that raters placed in the appraisal system influenced rater judgments, particularly leniency error; the ratings of those who reported low levels of trust in the performance appraisal system suffered from leniency error. Higher levels of leniency error have been found in organizations than in laboratory settings, a finding that is attributed to the use of performance appraisals to make high-stakes decisions in organizational settings. In contrast, ratings obtained in laboratory studies have few real consequences, and raters are less likely to be motivated to inflate ratings (Jawahar & Williams, 1997). Additionally, self-ratings are thought to be susceptible to leniency bias, although this bias may be mitigated when ratings are carefully reviewed by the rater's superiors (Bretz et al., 1992).

Halo error is the tendency for raters to give similar ratings across different and often distinct aspects of performance. Halo is thought to be a product of raters letting global evaluations influence the specialized evaluations for each dimension (Murphy, Jako, & Anhalt, 1993; Saal, Downey, & Lahey, 1980). Murphy et al. (1993) suggested that the traditional conceptualization of halo, in which it is considered to be rater error with negative influences, is misguided. Rather than representing an error specific to performance appraisal, halo probably reflects the basic cognitive processes raters follow when evaluating subordinates (Murphy et al., 2004). Indeed, no clear evidence has been found that raters are even capable of ignoring general evaluations when evaluating specific aspects of performance.

Laboratory studies have provided convincing evidence that the assumption that ratings that are free of halo or leniency are accurate assessments of performance is not correct. On the contrary, correlations between rater error measures and accuracy measures are generally small (Murphy & Balzer, 1989). In fact, evidence has shown that training that is designed to reduce rater errors such as leniency and halo can lead to lower levels of accuracy in ratings (Hedge & Kavanaugh, 1988). Rater error measures are slowly losing favor as criteria for evaluating ratings.

RELIABILITY AND CONSTRUCT VALIDITY OF PERFORMANCE RATINGS

The body of research on the reliability and validity of performance ratings is substantial; on the whole,

this literature paints a somewhat bleak picture. First, raters do not agree very well in their evaluations. Conway and Huffcutt (1997) examined reliability of supervisor, peer, and self-ratings. They found that although there was the highest consistency across supervisors, consistency was still disappointingly low, regardless of the source of evaluation. The most widely discussed reliability estimate was provided by Viswesvaran et al. (1996), who used interrater correlations to estimate reliability. They suggested that the reliability of performance ratings is approximately .52. Murphy and DeShon (2000) argued that interrater correlations do not provide acceptable reliability estimates and suggested applying generalizability theory to ratings. More recent estimates of the proportion of random error in performance appraisals (e.g., Hoffman, Lance, Bynum, & Gentry, 2010; Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Scullen, Mount, & Judge, 2003) have painted a less pessimistic picture of the reliability of performance ratings, but random measurement error is likely to account for at least 30% of the variance in performance ratings (Sturman & Murphy, 2010).

The somewhat limited reliability of performance ratings puts a limit on the level of validity that is to be expected in performance ratings. Although global assessments of construct validity have in some cases been quite favorable (e.g., a review of performance ratings conducted by the National Research Council [Milkovich & Wigdor, 1991] concluded that supervisory ratings of performance did show evidence of construct validity), more detailed assessments of validity have suggested that more caution may be needed in interpreting the meaning of ratings. For example, Scullen et al.'s (2003) confirmatory factor analysis suggested that lower order factors (technical skills, administrative skills, human skills, citizenship behaviors) had a good deal of construct validity but found that the higher order constructs were problematic (e.g., they hypothesized that ratings could be understood in terms of task vs. contextual performance).

On the whole, research has suggested that performance ratings are not as strongly related to job performance as most users of performance appraisal assume (Murphy, 2008). Researchers have differed substantially in their explanations of why the relationship between actual performance and performance ratings might be weak (e.g., Murphy, 2008, and Viswesvaran et al., 1996, fundamentally disagreed about what sources of variability should be treated as meaningful and what should be dealt with as error), but recognition is growing that the relationship between performance and performance ratings is a complex one and that broad statements about the validity of ratings may not be possible to make with much confidence. Valid and accurate ratings are probably most likely in environments in which raters are motivated to provide accurate assessments of performance and in which they have the tools and information to do so (Murphy & Cleveland, 1995). Unfortunately, performance ratings are often obtained under conditions in which raters have good reasons to distort ratings (e.g., because ratings will be used to distribute valued rewards) or in which they lack the knowledge and information needed to evaluate performance accurately (e.g., work groups in which supervisors have little opportunity to directly observe ratees). Despite our skepticism about the reliability, validity, and accuracy of the performance ratings that are often collected in organizations, there are reasons to be optimistic about the prospects for using performance ratings to accurately and honestly evaluate ratee performance in environments that actively and visibly support accuracy in rating.

A CLOSING NOTE: REASONS FOR OPTIMISM

Raters in organization often do not do a good job evaluating their subordinates. This common finding does not mean that they cannot do a good job evaluating performance. Rather, the shortcomings of performance appraisal are often a product of the environment in which ratings are obtained. Organizations that rely on the judgments of a single, poorly trained supervisor (with no meaningful cross-checks), obtained using vaguely worded appraisal forms, are sending a powerful message about the priority and value they assign to performance appraisals. Organizations that fail to recognize conscientious raters or that tolerate blatantly inaccurate ratings are reinforcing this message. One possibility is that raters who treat performance evaluation as a

nuisance and a joke are simply taking their cue from the organization.

Organizations are much more likely to obtain honest, accurate, and reliable ratings if they clearly and credibly value and support such ratings. In particular, organizations that tie rewards and sanctions to providing valid and useful performance ratings are much more likely to get good ratings than organizations that ignore the quality of ratings (Murphy & Cleveland, 1995). Organizations that hold raters accountable in a meaningful way (e.g., by conducting in-depth reviews of performance evaluations) are much more likely to end up with valid and useful ratings than organizations that clearly do not care about the quality of rating data. Finally, organizations that provide raters with the information, the tools, and the training to evaluate performance well are more likely to produce valid and useful ratings than organizations that leave the rater to his or her own devices.

Our most compelling basis for optimism is our experience working with organizations that take performance appraisal seriously. High-quality performance appraisal is possible, but it is a lot of work. Nevertheless, the shortcomings of performance appraisal in most organizations do not seem to be a function of the inability of raters to make good judgments or the basic intractability of the task of performance appraisal. Rather, most organizations get the sorts of performance appraisals they deserve. If an organization is willing to devote the time and resources needed to develop and maintain high-quality performance appraisal systems, they have good reasons to be optimistic about the ultimate quality of their appraisal systems.

References

Ashford, S. J., Blatt, R., & VandeWalle, D. (2003). Reflections on the looking glass: A review of research on feedback-seeking behavior in organizations. *Journal of Management, 29*, 773–799.

Atkins, P. W. B., Wood, R. E., & Rutgers, P. J. (2002). The effects of feedback format on dynamic decision making. *Organizational Behavior and Human Decision Processes, 88*, 587–604. doi:10.1016/S0749-5978(02)00002-X

Atwater, L. E., & Brett, J. F. (2006). Feedback format: Does it influence manager's reactions to feedback? *Journal of Occupational and Organizational Psychology, 79*, 517–532. doi:10.1348/096317905X58656

Atwater, L. E., Waldman, D. A., & Brett, J. F. (2002). Understanding and optimizing multisource feedback. *Human Resource Management, 41*, 193–208. doi:10.1002/hrm.10031

Austin, J. T., & Villanova, P. (1992). The criterion problem 1917–1992. *Journal of Applied Psychology, 77*, 836–874. doi:10.1037/0021-9010.77.6.836

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research practice gap in performance appraisal. *Personnel Psychology, 38*, 335–345. doi:10.1111/j.1744-6570.1985.tb00551.x

Barrett, C. V., & Kernan, M. C. (1987). Performance appraisal and termination: A review of court decisions since Brito v. Zia with implications for personnel practices. *Personnel Psychology, 40*, 489–503. doi:10.1111/j.1744-6570.1987.tb00611.x

Bernardin, H., Alvares, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology, 61*, 564–570. doi:10.1037/0021-9010.61.5.564

Bernardin, H. J. (1977). Behavioral expectation scales versus summated rating scales: A fairer comparison. *Journal of Applied Psychology, 62*, 422–427. doi:10.1037/0021-9010.62.4.422

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston, MA: Kent.

Bernardin, H. J., & Orban, J. A. (1990). Leniency effect as a function of rating format, purpose for appraisal, and rater individual differences. *Journal of Business and Psychology, 5*, 197–211. doi:10.1007/BF01014332

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *Journal of Applied Psychology, 66*, 458–463. doi:10.1037/0021-9010.66.4.458

Bjerke, D. C., Cleveland, J. N., Morrison, R. F., & Wilson, W. C. (1987). *Officer fitness report evaluation study* (Navy Personnel Research and Development Center Report TR 88–4). San Diego, CA: Navy Personnel Research and Developmental Center.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587–605. doi:10.1111/j.1744-6570.1995.tb01772.x

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 100–120). Baltimore, MD: Johns Hopkins University Press.

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*, 1–21. doi:10.1207/s15327043hup0601_1

Borman, W. C., Bryant, R. H., & Dorio, J. (2010). The measurement of task performance as criteria in selection research. In J. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 439–461). Mahwah, NJ: Erlbaum.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco, CA: Jossey-Bass.

Boswell, W. R., & Boudreau, J. W. (2000). Employee satisfaction with performance appraisals and appraisers: The role of perceived appraisal use. *Human Resource Development Quarterly, 11*, 283–299. doi:10.1002/1532-1096(200023)11:3<283::AID-HRDQ6>3.0.CO;2-3

Bracken, D., Timmreck, C., & Church, A. (2000). *Handbook of multisource feedback*. San Francisco, CA: Jossey-Bass.

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance research and practice: Concerns, directions, and implications. *Journal of Management, 18*, 321–352. doi:10.1177/014920639201800206

Campbell, J. P. (1990). An overview of the Army Selection and Classification Project. *Personnel Psychology, 43*, 231–239. doi:10.1111/j.1744-6570.1990.tb01556.x

Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for personnel decisions. *Personnel Psychology, 34*, 211–226. doi:10.1111/j.1744-6570.1981.tb00939.x

Cleveland, J. N., & Murphy, K. R. (1992). Analyzing performance appraisal as goal-directed behavior. In G. Ferris & K. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 10, pp. 121–185). Greenwich, CT: JAI Press.

Cleveland, J. N., Murphy, K. R., & Lim, A. (2007). Feedback phobia? Why employees do not want to give or receive it. In J. Langan-Fox, C. Cooper, & R. Klimoski (Eds.), *Research companion to the dysfunctional workplace: Management challenges and symptoms* (pp. 168–186). Cheltenham, England: Edward Elgar.

Cleveland, J. N., Murphy, K. R., & Williams, R. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130–135. doi:10.1037/0021-9010.74.1.130

Coen, T., & Jenkins, M. (2000). *Abolishing performance appraisals: Why they backfire and what to do instead.* New York, NY: Berrett-Koehler.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331–360. doi:10.1207/s15327043hup1004_2

DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review, 3*, 635–646.

DeNisi, A. S. (1996). *Cognitive processes in performance appraisal: A research agenda with implications for practice*. London, England: Routledge.

Dorfman, P. W., Stephan, W. G., & Loveland, J. (1986). Performance appraisal behaviors: Supervisor perceptions and subordinate reactions. *Personnel Psychology, 39*, 579–597. doi:10.1111/j.1744-6570.1986.tb00954.x

Dornbusch, S. M., & Scott, W. R. (1975). *Evaluation and the exercise of authority*. San Francisco, CA: Jossey-Bass.

Farh, J.-L., & Werbel, J. D. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. *Journal of Applied Psychology, 71*, 527–529. doi:10.1037/0021-9010.71.3.527

Farr, J. L., & Jacobs, R. (2006). Trust us: New perspectives on performance appraisal. In W. Bennett, D. Woehr, & C. Lance (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 321–337). Mahwah, NJ: Erlbaum.

Gioia, D. A., & Longenecker, C. O. (1994). Delving into the dark side: The politics of executive appraisal. *Organizational Dynamics, 22*(3), 47–58. doi:10.1016/0090-2616(94)90047-7

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*, 737–756. doi:10.1016/0149-2063(94)90028-0

Harris, M. M., Ispas, D., & Schmidt, G. (2008). Inaccurate performance appraisal ratings are a reflection of larger organizational issues. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 190–193. doi:10.1111/j.1754-9434.2008.00037.x

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisory, self-peer, and peer-supervisory ratings. *Personnel Psychology, 41*, 43–62. doi:10.1111/j.1744-6570.1988.tb00631.x

Hedge, J. W., & Kavanaugh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraisal training. *Journal of Applied Psychology, 73*, 68–73. doi:10.1037/0021-9010.73.1.68

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39*, 811–826. doi:10.1111/j.1744-6570.1986.tb00596.x

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119–151. doi:10.1111/j.1744-6570.2009.01164.x

Hollenbeck, G. P. (2008). When I use a word . . . *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 183–184. doi:10.1111/j.1754-9434.2008.00035.x

Hunt, S. T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology, 49*, 51–83. doi:10.1111/j.1744-6570.1996.tb01791.x

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology, 33*, 595–640. doi:10.1111/j.1744-6570.1980.tb00486.x

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*, 905–925. doi:10.1111/j.1744-6570.1997.tb01487.x

Kane, J. S. (1986). Performance distribution assessment. In R. Berk (Ed.), *The state of art in performance assessment* (pp. 237–273). Baltimore, MD: Johns Hopkins University Press.

King, J. (2008). How managers think: Why a mediated model makes sense. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 180–182. doi:10.1111/j.1754-9434.2008.00034.x

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. doi:10.1037/0033-2909.119.2.254

Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. *Applied Psychological Measurement, 3*, 193–200. doi:10.1177/014662167900300209

Landy, F. J., & Conte, J. M. (2007). *Work in the 21st century: An introduction to industrial and organizational psychology* (2nd ed.). Malden, MA: Blackwell.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107. doi:10.1037/0033-2909.87.1.72

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications.* New York, NY: Academic Press.

Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foremen. *Personnel Psychology, 32*, 299–311. doi:10.1111/j.1744-6570.1979.tb02136.x

LePine, J. A., Erez, A., & Johnson, D. E. (2002). The nature and dimensionality of organizational citizenship behavior: A critical review and meta-analysis. *Journal of Applied Psychology, 87*, 52–65. doi:10.1037/0021-9010.87.1.52

Leung, K., Su, S., & Morris, M. W. (2001). When is criticism not constructive? The role of fairness perceptions and dispositional attributions in employee acceptance of critical supervisory feedback. *Human Relations, 54*, 1155–1187. doi:10.1177/0018726701549002

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30*, 881–905. doi:10.1016/j.jm.2004.06.005

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*, 709–734.

McGregor, D. (1957). An uneasy look at performance appraisal. *Harvard Business Review, 35*, 89–94.

Meyer, H. H., Kay, E., & French, J. (1965). Split roles in performance appraisal. *Harvard Business Review, 43*, 123–129.

Milkovich, G. T., & Wigdor, A. K. (1991). *Pay for performance.* Washington, DC: National Academies Press.

Morales, M., & Bautista, R. (2008, August). *Planned missingness with multiple imputation: Enabling the use of exit polls to reduce measurement error in surveys.* Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater, and level effects in 360-degree performance rating. *Personnel Psychology, 51*, 557–576. doi:10.1111/j.1744-6570.1998.tb00251.x

Murphy, K. R. (1989). Dimensions of job performance. In R. Dillon & J. Pelligrino (Eds.), *Testing: Applied and theoretical perspectives* (pp. 218–247). New York, NY: Praeger.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 148–160. doi:10.1111/j.1754-9434.2008.00030.x

Murphy, K. R., & Balzer, W. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619–624. doi:10.1037/0021-9010.74.4.619

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational and goal-oriented perspectives.* Newbury Park, CA: Sage.

Murphy, K. R., Cleveland, J. N., & Mohler, C. (2001). Reliability, validity and meaningfulness of multisource ratings. In D. Bracken, C. Timmreck, & A. Church (Eds.), *Handbook of multisource feedback* (pp. 130–148). San Francisco, CA: Jossey-Bass.

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*, 158–164. doi:10.1037/0021-9010.89.1.158

Murphy, K. R., & Constans, J. L. (1987). Behavioral anchors as a source of bias in rating. *Journal of Applied Psychology, 72*, 573–577.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900. doi:10.1111/j.1744-6570.2000.tb02421.x

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). The nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*, 218–225. doi:10.1037/0021-9010.78.2.218

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology, 67*, 562–567. doi:10.1037/0021-9010.67.5.562

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2008). No new terrain: Reliability and construct validity of job performance ratings. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 174–179. doi:10.1111/j.1754-9434.2008.00033.x

Organ, D. W. (1988). *Organizational behavior: The good soldier syndrome*. Lexington, MA: Lexington.

Patz, A. L. (1975). Performance appraisal: Useful but still resisted. *Harvard Business Review, 53*, 74–80.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, D. G. (2000). Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management, 26*, 513–563. doi:10.1177/014920630002600307

Roch, S. G., Sturnburgh, A. M., & Caputo, P. M. (2007). Absolute vs. relative rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15*, 302–316. doi:10.1111/j.1468-2389.2007.00390.x

Saal, F. E., Downey, R. C., & Lahey, M. A. (1980). Rating the ratings: Assessing the quality of rating data. *Psychological Bulletin*, *88,* 413–428. doi:10.1037/0033-2909.88.2.413

Scullen, S. E., Bergey, P. K., & Aimon-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology*, *58*, 1–32. doi:10.1111/j.1744-6570.2005.00361.x

Scullen, S. E., Mount, M. K., & Judge, T. A. (2003). Evidence of the construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology, 88*, 50–66. doi:10.1037/0021-9010.88.1.50

Sims, H. P., Jr., Gioia, D. A., & Longenecker, C. O. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1*, 183–193. doi:10.5465/AME.1987.4275731

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149–155. doi:10.1037/h0047060

Smither, J. W., & Walker, A. G. (2004). Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time? *Journal of Applied Psychology, 89*, 575–581. doi:10.1037/0021-9010.89.3.575

Sturman, M., & Murphy, K. (2010, August). *Sources of error variance and their effects on supervisor's job performance ratings*. Paper presented at the annual conference of the Academy of Management, Montreal, Quebec, Canada.

Thornton, G. C., III. (1980). Psychometric properties of self-appraisals of job performance. *Personnel Psychology, 33*, 263–271. doi:10.1111/j.1744-6570.1980.tb02348.x

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Erlbaum.

Tziner, A., & Murphy, K. R. (1999). Additional evidence of attitudinal influences in performance appraisal. *Journal of Business and Psychology, 13*, 407–419. doi:10.1023/A:1022982501606

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2001). Relationships between attitudes toward organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment, 9*, 226–239. doi:10.1111/1468-2389.00176

Tziner, A., Murphy, K. R., Cleveland, J. N., Beaudin, G., & Marchand, S. (1998). Impact of rater beliefs regarding performance appraisal and its organizational contexts on appraisal quality. *Journal of Business and Psychology, 12*, 457–467. doi:10.1023/A:1025003106150

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574. doi:10.1037/0021-9010.81.5.557

Welch, J. F. (2001). *Jack: Straight from the gut*. New York, NY: Warner Books.

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17*, 601–617. doi:10.1177/014920639101700305

CHAPTER 34

IMPLEMENTING ORGANIZATIONAL SURVEYS

Paul M. Connolly

An organization functions only as well as its people function in their jobs and with each other in the pursuit of business goals. There are many different approaches to the assessment of organizational functioning, or identifying the ways in which people work effectively with each other to achieve the organization's goals (Church, Waclawski, & Kraut, 2001; Edwards, Thomas, Rosenfeld, & Booth-Kewley, 1996; Wiley, 2010). This chapter is built on the author's perspective and experience with hundreds of organizations using three types of organizational surveys to perform an assessment (Connolly & Connolly, 2005).

WHY ASSESS ORGANIZATIONAL FUNCTIONING?

Assessments in themselves do not increase organizational or individual excellence. Rather, they create an understanding of issues and practices that help or hinder individual performance. Once issues are identified, survey results can provide a platform for changes and improvements. Here are three ways in which assessments help increase organizational functioning:

1. The announcement of a group assessment communicates that the sponsor wants data and will not form opinions unilaterally. It demonstrates organizational commitment to listening. Participative decision making has demonstrated many important outcomes, including positive effects on employee judgments of fairness (Witt, Andrews, & Kacmar, 2000).
2. Effective assessments require the use of a common language, shared processes, schedules, and plans for postassessment activities. As a result, assessment builds consensus on basic management processes that positively influence all functions of the organization. It sends the message "We are all working together to improve the organization."
3. Assessment typically leads to metrics, a key part of a framework for measuring gaps, deficits, and strengths and for describing and measuring improvement goals (Gallup Organization, 1998).

Yet, because assessments are not an end in themselves, this simple maxim is a worthy goal: The objective of a survey is to make the need for another survey go away. If this objective is reached, people become comfortable identifying, discussing, and resolving business-related problems. The role of formal assessment diminishes. The issues get resolved as they come up, without the need for an assessment.

The goals of this chapter are threefold. First, the chapter provides a solid framework for understanding the types of organizational assessments used for both groups and individuals. It also considers how assessments interrelate. Second, the chapter aims to help organizational leaders become aware that the assessment process itself does not increase organizational functioning, but it does create common rules

The author thanks long-time coauthor (and spouse) Kathleen Groll Connolly for her support and guidance in preparing this chapter.

DOI: 10.1037/14047-034
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

of engagement that can and do help. It also creates a data-rich asset for measurement and comparison as strategies unfold and business conditions change. Both individual and group assessment can create a common ground for inquiry, conversation, action, and closure. When the rules are thoughtfully defined and the language is reliably applied, the assessment process can work like magic to increase organizational functioning. Third, the chapter offers a broad overview of organizational assessment implementation.

A MODEL OF ASSESSMENT

Why should an organization make a commitment to assessment and metrics? Quite simply, metrics—both individual and organizational—can bring a level of rationality to the dynamic and unpredictable process of running an organization.

Once leaders commit to metrics as part of their performance strategy, however, many other questions remain to be answered. How are assessments selected? Given an organizational situation, should individuals, the entire organization, or both be assessed? In what sequence should individuals and organizations be studied? Indeed, individual assessment, when it is conducted with many members of the organization, can have an impact similar to that of assessing the organization as a whole (Hogan, 2007). Provided with the opportunity for their own assessments, individuals often increase their commitment to the employer (Schiemann, 1996). Likewise, provided with the opportunity to give opinions about organizational functioning, many individuals also increase their commitment to the employer (Hinrichs, 1996).

Yet, individual assessments and organizational assessments are very different tools. Individual assessments can help people improve performance, whereas organizational assessments help leaders sculpt a culture in which individuals can function together. Chapters 24 through 30 in this volume cover individual assessments in great detail, and it is useful to discuss their link with organizational assessment. Figure 34.1 shows a simple way to conceptualize this link.

Organizations are essentially collections of individuals who bring their intellect, personalities, motives, and behaviors to work. They learn, grow, and change in their experience within the organization. The group forms the boundary in which people find (or do not find) their fit.

Four categories of individual assessments are well-established today, as shown in Figure 34.1. Assessments may deal with cognitive capacity (how smart), personality (what tendencies),

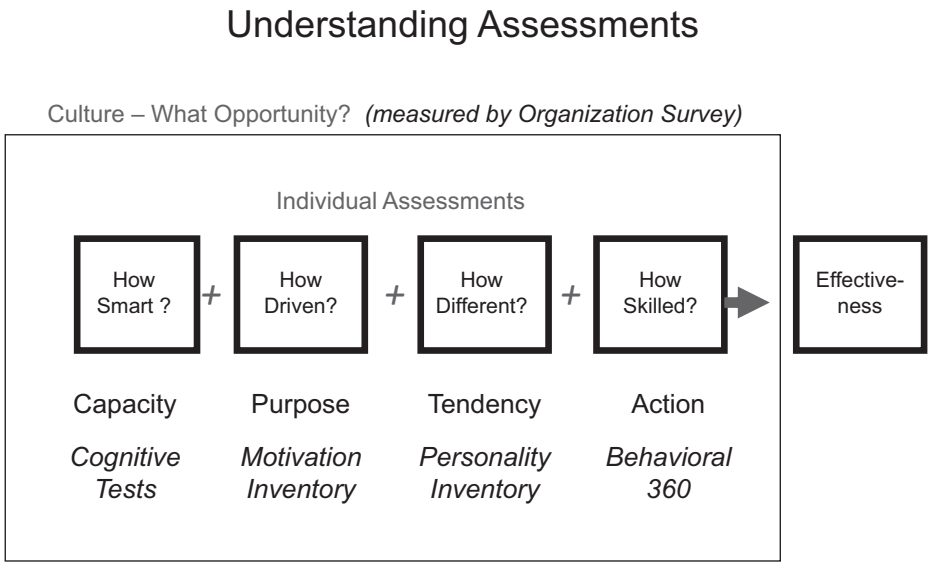

FIGURE 34.1. A model for understanding assessments. Copyright 2011 by Performance Programs, Inc. Used with permission.

motivation (what purpose), and behavior (what actions). As organizations respond to marketplace opportunity, they may use these assessments to select, promote, or train individuals, and these efforts do have an impact on the organizational culture.

On the other hand, organizational culture is the context in which all these individual differences exist, and it represents the field of opportunity in which individuals can have an impact. Many organizations focus a great deal of attention on selecting and developing people. Perhaps the clearest link between individual and organizational assessment is *culture fit,* which is basically the connection between a person's values and the organization's culture. Problems can arise quickly when there is a clash between individual motivation (purpose, or what an individual wants to achieve) and organizational culture (which influences what types of behavior are rewarded; Herzberg, 1968; Hogan, 2007). The realities of individual differences are acknowledged, but this chapter focuses on how to assess—and in the process improve—the organization's effectiveness.

There are three basic approaches to assessing an organization: audit or fact-finding surveys, alignment surveys, and engagement surveys.

Audit or Fact-Finding Surveys

When organizations first embark on assessment, they commonly focus on fact finding. The goal of audit surveys is to gather baseline information for use by management in developing metrics. This first audit is often conducted through conversations with management, review of cross-organizational performance data, focus groups, or pretesting survey concepts with small segments of the organization's population. Many audits are topic focused, seeking information on some specific aspect of organizational life. Often, the results of these audits are not shared with the people who supplied the information (Kraut, 1996). Perhaps the most famous example of this approach is the survey program run by Sears beginning in 1938 (Dunham & Smith, 1979). For many years, Sears collected information from employees to guide management practices.

Alignment Surveys

Think of alignment as "walking the talk." In assessments of alignment, one looks for differences between stated visions, values, and goals and the actual practices carried out by different facets of an organization. The goal of an alignment survey is to test how well a vision, value, or belief has been communicated and embraced by an organization (Schneider, Salvaggio, & Subirats, 2002). The evidence that these surveys can cause change and help transform workplace culture is clear (Wagner & Spencer, 1996).

Typically, alignment surveys use formal data-gathering methods and involve the entire population. Information is summarized by groups, and the summary is reported back to those groups for discussion and training. Unlike most audits, these surveys involve two-way communication.

To understand how an organization can benefit from an assessment of alignment, consider the case of a religiously based health care organization that had a strong statement of values (justice, charity, respect, etc.) and spent a great deal of time clarifying what that statement meant for employees in their work with patients. However, the organization was encountering conflict around issues of respect among coworkers. The management suspected that it had neglected to emphasize the importance of the core values for employees working with one another. An alignment survey assessed the extent to which employees believed those practices were in place, with items such as "I feel I am treated with dignity and respect by my coworkers" and "My immediate manager is supportive of my efforts to get the job done."

The survey revealed gaps between the stated values and actual practices and caused management to refine job descriptions and offer training. It was also followed by meetings in which results were shared. Steps were taken to ensure that employees felt safe reporting impediments to carrying out the values.

Engagement Surveys

In the 1990s, engagement assessment was becoming the most common survey approach used by organizations (Higgs & Ashworth, 1996), and in my experience that trend continues today. The goal of engagement surveys is to involve employees in both

problem identification and resolution. Engaged employees not only point out problems but work toward resolving them. Hinrichs (1996) cited several studies demonstrating positive organizational changes via engagement surveys and the feedback process, as did Davenport, Harris, and Shapiro (2010). Engagement surveys are also called *opinion, viewpoint, morale, satisfaction, involvement,* or *commitment* surveys. Although distinctions can be made among these survey types, they are all basically assessments driving toward the same goal: productive workplace environments.

When engagement is the purpose, the assessment process is most often designed to involve employees in finding solutions to issues hampering organizational performance. One such case was a regional health insurance company looking to compete effectively for both employees and customers. It held focus groups with employees to determine the areas that a survey should cover. The company designed the survey to ensure that results would be meaningful to relatively small work units, so those groups could meet to discuss survey results and possible solutions for any problem areas that were identified. Over the course of 10 years and five survey cycles, the response rate went from 78% to 93% participation, and the number of serious long-term problems identified by the survey went from 10 to three.

IMPLEMENTATION OF ORGANIZATIONAL ASSESSMENTS

Effective organizational assessment involves as many as six major steps. Audit surveys require the first five steps, and alignment and engagement surveys include the sixth step. Exhibit 34.1 summarizes these steps.

Step 1: Planning

A thorough, clear plan is essential to the success of any survey program. The plan needs to include how all employees will be involved, how information will be collected, who will be surveyed and when, and what steps will be taken to assure a high response rate. We look at each of these elements in detail in the sections that follow.

Exhibit 34.1
The Organizational Assessment Process

Step 1: Planning & communication
Step 2: Assessment content
Step 3: Information collection
Step 4: Information reporting
Step 5: Analysis and interpretation
Step 6: Feedback & action planning

Note. Copyright 2011 by Performance Programs, Inc. Used with permission.

Initial structure. Once the need for an audit, alignment, or engagement survey has been established, the first action is to identify a sponsor and a project manager. The sponsor should be the highest organizational official willing to support the effort, who will very often be the manager requesting the survey. The sponsor generally appoints two other people or groups at this stage. One of these is the project manager, often someone from the human resources department. The project manager develops an initial action plan, which often includes the creation of a steering team.

The steering team is usually a small group consisting of those who will support implementation. In larger organizations, the steering team often has people from specialties such as internal communications and information technology whose technical skills are required to get the job done. In addition, representatives of departments or divisions that will be surveyed generally join the steering team. In smaller organizations, the steering team may consist of only one or two people. This team might include, for instance, a human resources manager and an executive assistant to the president. Either way, the sponsor or the sponsor's representative should always participate in the steering team as well.

Gathering input. Typically, alignment or engagement surveys begin with an audit or fact-finding stage. As mentioned earlier, this stage may be carried out by means of focus groups or interviews. Both of these approaches can provide valuable input into structuring the assessment. Typically, these groups meet with a facilitator and often provide guidance

on survey content. They may also make suggestions about the process, including rollout, execution, reporting, feedback, and follow-up.

Defining the survey population. Most surveys today involve the entire employee population, for two reasons. First, online survey platforms limit the cost of inclusion to the cost of the employee's time. Including one more person results in no additional paper, copying, mailing cost, or data entry. Second, particularly if the survey objective is engagement, it helps to involve as many people as possible as early as possible. Asking for survey participation actually supports engagement.

There are times when organizations still use sampling, however. Audits, which are usually only intended to achieve a broad identification of issues, often involve only a portion of the employee population. Table 34.1 gives an idea of how few can be sampled (second column) if the organization wants to represent the views of the many (first column) with a sampling accuracy of ±5% (Connolly & Connolly, 2006). Such sampling tables can be found in many statistics texts, including Cochran (1963).

TABLE 34.1

Representative Sample Sizes

When the group you want to represent is this large:	Then you need this many responses:
10	10
25	24
40	36
55	48
70	59
85	70
100	80
200	132
300	169
400	196
500	217
600	234
700	248
800	265
900	274
1,000	285
2,000	322
3,000	341
4,000	351
5,000	357
6,000	361
7,000	364
10,000	370

Note. Copyright 2011 by Performance Programs, Inc. Used with permission.

Determining the timing. A survey is always a snapshot in time, and the timing selected can be important. A focus group might help choose a time with the fewest disadvantages. A word of caution about timing surveys: Think carefully about doing a survey before any major organizational change, such as a layoff, a corporate sale, or during a union organizing campaign (when surveying could actually be deemed illegal). As long as employees are aware of the impending change, problems will generally not occur. In fact, survey results can provide a useful barometer for the success of the change. If the upcoming change is not known, however, employees may later become suspicious that their responses will be (or have been) used to help structure the surprise decision. Whether that is true or not, it may take many years before employees will trust a survey process again. A 2-week response window is recommended, with a 3rd week optional if returns are disappointing near the end of the 2nd week. During holiday periods, allow 3 weeks with an optional 4th week.

Ensuring good response rates. Communication is everything in obtaining responses, but that communication has to be preceded by some careful policy decisions regarding confidentiality and anonymity. Anonymity and confidentiality have a subtle but critical difference between them. Those concerned with anonymity ask, "Will my answers be identified as mine by anyone, in any way?" When people believe their responses are anonymous, they are more likely to say what is on their mind without fear of reprisal. For example, one company took steps to create the conditions for truly anonymous response, which resulted in employees identifying a harassment situation that was in violation of company policy. The situation, once identified, could be investigated and remedied. When an organization is seeking information about possible violations, complete anonymity is highly recommended. Sudman and Bradburn (1974) as well as Ong and Weiss (2000)

have provided a good discussion of anonymity and truthfulness of responses, especially to sensitive questions.

A promise of confidentiality is not, however, necessarily a promise of anonymity. Confidential surveys deliver results only to specified parties, who respect the respondents' privacy even as they deliver essential information from survey results. *Confidentiality* also refers to the idea of a survey result being delivered only to those who need to know, not necessarily to everyone in an organization. For example, if a well-liked manager had developed a bad personal habit, an individual could use the survey to comment on it, knowing that the results would only be shared internally. This avoids embarrassing the manager outside of the work group. It can be helpful to announce beforehand that work group results will not be reported across the organization. Those results might be shared with division or senior management, but not with peer work groups. The underlying issue is one of trust and one that needs to be communicated at each step of the survey process.

Successful survey communications have to include several key points in addition to anonymity and confidentiality. They must explain the value of the information to the organization's future. (It helps if this explanation comes from someone to whom employees feel a strong leadership connection.) Communications must also clarify the roles to be played by various individuals and groups during survey implementation as well as upcoming steps. Many sources of sample communications letters, announcements, and materials can be found online and in many "how-to" survey books.

Step 2: Assessment Content

The planning process determines the assessment's breadth of coverage. Then survey questions must be selected from lists of standard items or created.

Standard and custom questions. Standard questions are prewritten and, one hopes, pretested. Such questions might be obtained from a prior survey, from a book or online resource containing survey items, or from a survey vendor. Standard questionnaire items usually refer to issues affecting all organizations, so they tend to be generic. Also, because they have been used before, they are likely through trial and error to be well worded and thus more likely to provide useful information.

It takes time to write good survey items, so the use of standard questions can be a considerable time saver (Sudman & Bradburn, 1982). One note of importance: Standard questions can be slightly rephrased without influencing the statistics related to them. For example, changing *company* to *organization* is not worth worrying about, nor would it likely be a problem to change *employee* to *associate,* if that is how people within the organization are addressed (Schaeffer & Presser, 2003).

Customized questions, by contrast, are more likely to focus on specific and unique issues within the organization. For example, one might want to ask a specific question about an employee publication or specific benefit program. It may take several rounds of writing, testing, and rewriting to create a useful item. Most surveys include both standard and custom questions in an effort to be relevant and also to have external comparisons.

Rating scales. Survey scales use everything from simple yes–no questions to 3-point, 5-point, 7-point, or even 10-point ratings. Research has indicated that using a 5- to 9-point scale is optimal (Miller, 1956). Fewer than five rating choices may make respondents feel constrained; more than nine options seems overly complex. The 5-point Likert scale is that most commonly used in organizational surveys. Usually the scale ranges from the lowest rating to the highest rating, although some prefer the reverse. In practice, direction makes less difference than consistency of direction.

The most common option is the agreement scale that ranges from *strongly disagree* to *strongly agree.* Other scales can be used for frequency (*never* to *always*), satisfaction (*very dissatisfied* to *very satisfied*), amount (*none* to *all*), or importance (*low* to *high*).

Some prefer an even number of rating points (no middle rating) instead of an odd number, usually because of a desire to force people to take one side or the other of an issue. Our data have shown that a 4-point scale usually results in forcing people who are neutral to inflate their response in a favorable direction, a finding consistent with that reported by

Bishop (1987). In other words, even-point scales add error, which is why most surveys should have an odd number of rating points.

Number of questions. Thirty years ago, an employee survey not uncommonly had 200 or more questions and took up to an hour to complete. Today, the acceptable window of time is often 20 minutes. With modern technology, 20 minutes translates to about 70 questions. Pulse surveys, which are short topical surveys, often have 10 questions or fewer.

Survey fatigue is counteracted by a few different factors. Clustering similar items, using a consistent rating scale, using a scale with a content-appropriate number of options, using simple terms, avoiding too-similar items, providing an easy-to-use web interface—all of these things work to reduce fatigue and support survey participation.

Open-ended or written comments. All organizational surveys should provide the opportunity for comment. Comments provide insight into the emotions and feelings behind the numerical results as well as perspective on issues not captured by the rated items. They also communicate the willingness of the survey sponsor to receive answers to questions that were not asked but are on the respondents' minds. Comments sections can be provided after each set of items, as discussed earlier, or at the end of the survey.

Demographics. Most organizational assessments include demographics that facilitate analysis by subgroupings of interest to management. The most commonly used demographic for organizational surveys is work group or department because, particularly for engagement surveys, the key advantage of the assessment comes from having small group meetings to discuss the group's results.

Another common demographic in recent surveys has been age group, especially as it relates to generational differences. Lancaster and Stillman (2002) suggested age ranges to separate viewpoints by generation. With respect to communication, traditionalists (born between 1900 and 1945) often do not share information up or down levels. Baby boomers (born between 1946 and 1964) prefer annual formal communication and feedback, with weekly informal discussion. Generation Xers (born between 1965 and 1980) give and seek immediate and regular feedback at all times. Millennials, or Gen Ys (born between 1981 and 1999), prefer visual, instant electronic communication. It is hardly surprising that these cohorts would have differing views of organizational life, so age group has become a more common survey demographic.

Some organizations use the annual or semiannual survey as a way to monitor their affirmative action programs, so they may ask respondents to indicate their race or gender. When an organization does not want to seek sensitive information outright, one strategy is to ask the respondent whether he or she considers him- or herself "a member of a minority group based on race, religion, gender or sexual orientation."

Other common demographics look at organizational level (senior executive, manager, supervisor, etc.). Tenure is often skipped because the results are typically U-shaped, that is, those who are new are usually positive, those who are less new are less positive, and those with longer tenure are more positive (i.e., short termers are generally optimistic and unhappy people leave).

For survey respondents, demographic questions can be a red flag. The more demographic information a survey requests, the greater the skepticism about survey anonymity. In a 100-person company, how many female Asian/Pacific Islanders with 15 or more years of service between the ages of 45 and 55 are there? Too many demographics can undermine trust. Focus groups may also guide decisions about the acceptable number of demographic questions.

Jargon, idiom, and multiple languages. Even without translation, some terms will cause trouble in some industries. An example of this is using the word *safety* or *risk* in a manufacturing setting versus, say, an entrepreneurial investment start-up. This type of problem is often caught in the focus-group or pilot-test phase of the assessment process (Volume 3, Chapter 26, this handbook, provides considerably more information on this topic).

Multinational and multilingual surveys need a great deal of care in planning and execution.

Communications as well as logistics are likely to require more time. Avoid common jargon or slang, both of which tend to be difficult to translate. If possible, avoid industry jargon. Also, brevity of questionnaire items reduces the opportunity for confusing translations.

Step 3: Information Collection

Information collection has changed dramatically and mostly for the better with the expansion of the Internet. In the author's recent experience, 95% of organizations do their surveys online; however, some still use the traditional paper-and-pencil method, which actually still provides the highest response rates in places in which employees are all in one location, the company culture favors face-to-face meetings, or computers are not part of employees' daily work experience.

Many good online survey platforms are available, and some of them will even permit short surveys free or at very low cost. Three common problems arise with online surveys: access, time-outs, and spam blockers. Regarding access, some organizations give different levels of web access to different groups of employees. If someone is invited to participate in a survey, the part of the organization that sponsors the survey has an obligation to contact the information technology group and request the removal of blocks to access.

Regarding time-outs, information technology departments often set an amount of time that employees can remain on a Web page before automatically signing them out. This is done to manage bandwidth resources. In practical terms, if an employee starts the survey, goes to lunch, and begins again, chances are nearly certain in most organizations that when the employee hits "complete," he or she will get an error message. The employee may have to start over. Therefore, a special 12-hour window is highly recommended during survey administration.

Spam and e-mail blocking are becoming bigger problems. In some cases, information technology departments have been forced to restrict mass survey invitations. It is important to involve the information technology department early, obtaining agreement on a strategy for such individualized links.

Online access has implications for confidentiality and anonymity. A universal URL link to a survey has the advantage of simplicity and consistency. It is likely to create greater confidence that the organization is not tracking the responses of a specific person. The downside of the single link, however, is that someone can respond to the survey multiple times. There is no control on who can use the link. It can even be forwarded to those outside the organization, such as former employees.

One antidote to this problem is one-time-use passwords. This approach limits the number of times a person can respond, but it can leave the person feeling that responses can be identified with the password and undermine the claim of anonymity, even if anonymity is actually well protected. If the one-time-use password approach is used, be sure to communicate ahead of time why this approach is being used, and take steps to make sure passwords are not tied to survey results. For example, passwords should not be tied to the data record or responses. One strategy is to encourage people to change their one-time-use passwords. Another is to have them go online to a separate, third-party site that generates the password for them.

Step 4: Results—Reporting and Analysis

What mix of data and graphics is best? A mixture of both is usually best, with numbers for data-oriented people and graphics for picture-oriented people. Figure 34.2 shows the format used by the author of this chapter, which is very typical for the presentation of survey results.

For number- or data-oriented people, display the frequency distribution for each item (e.g., five people or 1% answered *strongly disagree*). Data people will also want to see the mean, or average, and the standard deviation, one measure of disagreement across raters.

For people who rely on graphics, it may be productive to combine results from each scale category into three categories: unfavorable, neutral, and favorable. This approach usually results in a quick transfer of information. Another useful section summarizes the scores of related items (scales) using the same data and graphic presentations. All "immediate manager" items, for instance, are averaged together

READING YOUR RESULTS

This report was designed to summarize the opinions given by individuals regarding the job and work environment at your organization. The information below describes how to read the results.

"PERSPECTIVE" refers to the group of employees rating each question.

"ACTUAL" refers to the actual number of individuals who responded to a particular question.

"NR" refers to the number of employees who left the question blank.

"% of RESPONSES" refers to the percentage of "Actuals" who responded with a 1, 2, 3, 4, & 5. Due to rounding, the totals of these percentages may be slightly above or below 100%.

"SD" refers to the standard deviation. SDs greater than 1.0 indicate a relatively high level of disagreement and should be examined more closely.

"AVG" refers to the average. This is the average of all responses for a particular question.

Bar Charts are based on the following scale:

"1" means *Strongly Disagree*.
"2" means *Disagree.*
"3" means *Neutral.*
"4" means *Agree.*
"5" means *Strongly Agree.*

The actual results from question 1 are given below as an example:

1. Overall, I am satisfied working for this Company at the present time.

			% Of Responses						Unfavorable -- Favorable
Perspective	**Actual**	**NR**	**1**	**2**	**3**	**4**	**5**	**Avg**	
All Sample Company	457	0	2	7	8	47	37	4.11	8 | 8 | 84
Central Division	82	0	0	4	6	44	46	4.33	4 | 6 | 90
Operations	50	0	0	0	8	40	52	4.44	8 | 92

Note: Ratings are indicated as a percentage value and graphically represented by a Bar chart.

UNFAVORABLE RATINGS (1's or 2's) are shown as a percentage and represented as a BLACK BAR on the accompanying chart.

NEUTRAL RATINGS (3's) are shown as a percentage and represented as a WHITE BAR on the accompanying chart.

FAVORABLE RATINGS (4's or 5's) are shown as a percentage and represented as a GRAY BAR on the

Depending on the actual results, there may be any combination of these three bars on your reports. The PREDOMINANT VIEW of each group is represented by the LONGEST BAR. Due to rounding, the combined percentages of these three bars may be slightly above or below 100%.

FIGURE 34.2. Sample employee survey report page. Copyright 2011 by Performance Programs, Inc. Used with permission.

to give a larger view. Both methods can be productively combined with a ranked summary of items—typically the "10 most favorable" and "10 least favorable"—to help people get a quick read on strengths and challenges.

Most organizations like a multiplicity of reports. The idea is to provide relevant information to people at all levels, avoiding either too much or too little information. As an example, a senior finance executive might have a four-line report. One line shows information for the entire organization, one line shows information for the finance department, another line shows a summary of responses for the six people in the audit department, and a final line shows information for the eight people in accounting. A senior human resources person might have a two-line report, one showing the responses for the entire organization and one showing information for the six people in human resources. Obviously the human resources report would exclude the finance information, and vice versa. It is valuable to have small groups see how they scored their work group or department compared with some other relevant groups. That practice helps focus the upcoming interpretation and review sessions, the subject of Step 5.

Step 5: Analysis and Interpretation

"How accurate are the results?" This is a natural first question about any survey report, but only one of a number of appropriate questions that lead to good analysis and interpretation. Accuracy—which can be thought of as how well the data represent reality—depends on a number of factors. One answer to the question, though, is found through the response rate—the number of people responding divided by the total number in the population who were eligible to respond. Once a survey is complete, a confidence interval can be estimated, based in part on the response rate.

Standards of interpretation. Four common standards are applied during survey interpretation: personal, comparative, external, and absolute.

The personal standard can be generated by simply considering whether the results are what a manager or managers would like them to be. The comparative standard uses some internal benchmark for comparison, such as prior survey results or department results versus whole organization results. The external standard compares results from one's organization with results from other organizations. This approach is also referred to as a *normative standard.* Norms are often available from trade or industry associations or from some consulting firms.

The absolute standard is the most widely used and consists of comparing means or percentages to a specific standard. A very commonly used absolute standard is based on the percentage of favorable and unfavorable results. It provides the basis for comparison shown in Table 34.2.

Neutral ratings. When responses cluster in the middle of a scale (say at 3 on a 5-point scale), interpretation can be puzzling. Some practitioners believe that if neutral ratings are 50% or more, the wording of the questionnaire item may be a problem (i.e., raters are not sure what the item is asking). The topic itself may be unclear or unimportant.

A more interesting possibility, however, is that opinions are in transition from negative to positive or vice versa. Any item with many neutral responses should be reviewed in feedback groups, which may provide some clarification. Unfortunately, neutrality may not be fully understood until a subsequent survey cycle.

TABLE 34.2

Survey Interpretation Rules of Thumb

Results	Interpretation
Favorable results (%)	
≥75	Outstanding strength
67–74	Strength
Unfavorable results (%)	
≥35	A critical issue
20–34	a danger zone
Neutral results (%)	
≥33%	Potential problem with item wording; potential positive to negative (or vice versa) trend occurring

Note. These standards of interpretation for favorable and unfavorable survey results are rules of thumb widely used by organizational consultants. This is one of several ways to interpret the meaning of survey outcomes. Copyright 2011 by Performance Programs, Inc. Used with permission.

Response patterns. Patterns among responses are generally more important than the response to any single questionnaire item. A quick way to get a glimpse of patterns is to generate the 10 most favorable ratings and the 10 most unfavorable ratings. Be alert for inconsistencies. For example, consider a small work group's results on two survey items:

1. *"My immediate manager gives me the support I need to do my job."* This item received 70% favorable ratings, placing it on the 10-most-favorable list.
2. *"Senior management provides the information I need to do my job."* This item received 45% unfavorable ratings, placing it on the 10-least-favorable list.

People felt supported by the immediate manager but did not feel they got as much support from senior management. This pattern flags a level of alienation or distance that cannot be positive for relations with senior management.

Interpreting written comments. Written comments often provide more negative or positive contrast than numerical results would lead one to expect. Comments express the feelings behind the numbers. There is a difference between viewing an average of 1.8 on a goal-setting item and an employee's comment that "I haven't a clue about what we are supposed to be doing."

One of the simplest ways to analyze comments is to simply split them into favorable and unfavorable categories. An additional refinement is to categorize them again by topic. Another approach is to assign a "heat index" to comments, with higher ratings assigned to irate comments and lower values assigned to more neutral suggestions.

It is advisable to have someone, perhaps a third party who is familiar with the organization, review comments for appropriateness before they are seen by management. Did someone inadvertently identify themselves in a comment? In cases such as these, add a notation "Comment edited to protect anonymity" and then make minimal edits that preserve the intent of the comment but provide reasonable protections to the respondent. Third-party review can also help flag instances in which serious charges are leveled at someone else.

Computer analysis of open-ended comments is an emerging field. Software products that can help interpret comments and methods to cull insights from these analyses are improving. Capabilities such as these are likely to encourage shorter surveys with more verbal content in the future.

Nonresponders. The only thing a researcher can know about nonresponders is that they did not respond. Nonresponse is most serious in audits, which typically do not have follow-up meetings as part of the process. If there are no feedback meetings after one of these surveys, there is no opportunity to hear from nonresponders.

For alignment and engagement surveys, nonresponders will have another opportunity for input at the time of the feedback meeting. If there is a high rate of nonresponse, one strategy is to try again with a simple, anonymous two-question survey:

1. Did you respond to the original survey?
2. If not, why not?

The reasons for nonresponse have often fallen into one of two categories. Either people felt the survey was a waste of time, which is often the case when a prior survey yielded no apparent action afterward, or people felt threatened, such as in workplaces where prior similar efforts coincided with negative consequences, such as layoffs.

Step 6: Feedback and Action Planning

If the organization is conducting an audit survey, its work is complete at the end of Step 5. If the organization is conducting an alignment or engagement survey, it will continue to Step 6. Well-run feedback meetings mobilize people to change (Hinrichs, 1996). Feedback is one way of thanking participants for the power they have given the organization by sharing information. It allows people to see that their views were recorded, and it also allows them to compare those views to others'. Small group meetings allow for discussion of results and clarification of needed actions.

How to do feedback. Whenever possible, results should be generated for individual work groups. These reports drive work group feedback sessions in which employees have an opportunity to learn

results and share ideas for improvement. The manager has a critical role at this stage, becoming the face of the survey process, so providing training and support on how to run a feedback meeting is worthwhile. Exhibit 34.2 provides a typical feedback meeting agenda (Connolly & Connolly, 2006).

Most organizations set aside 3 to 4 hours for two separate meetings. The purpose of the first meeting is to review results and clarify issues. The purpose of the second meeting is to identify which issues to address and some ideas to resolve issues, called *action plans.*

Action planning. We suggest encouraging work groups to select two issues from those revealed at the first meeting. The "rule of two" says the work group should pick one issue that it can resolve itself without outside input or approval and one issue that is outside of its control to change. This one simple idea achieves several objectives. First, it focuses action on a manageable number of changes. Second, it demonstrates that some of the changes can be achieved by the group itself. Third, it avoids dumping problems on others. It is much better to select two issues to resolve, resolve them, and then hold a follow-up action meeting 6 months later to select an additional item. Exhibit 34.3 shows a typical action plan format (Connolly & Connolly, 2006).

Exhibit 34.2
Sample Agenda for an Employee Feedback Meeting

1. Review survey purpose, format, and administration.
2. Review survey response rate.
3. Distribute results; provide quiet time for reading.
4. Review results.
5. Manager interprets results.
6. Conduct initial feedback session for reactions to manager's interpretation and suggested action plans.
7. Schedule follow-up meetings to determine progress on Action Plans from Step 6 and to develop more action plans.
8. Schedule action planning meetings.

Note. Copyright 2011 by Performance Programs, Inc. Used with permission.

Exhibit 34.3
Action Planning Issues Worksheet

WORKGROUP: DATE:
INSTRUCTIONS: Fill out one copy of this form for each topic discussed during the feedback meeting that represents a problem that cannot be resolved by your group. Select ONE of these issues as your external priority, and forward a copy to Human Resources/ Survey Action Committee.
ISSUE SUMMARY:
ACTION RECOMMENDATIONS:
PERSON RESPONSIBLE FOR FOLLOW-UP (IN YOUR UNIT):
TARGET COMPLETION DATE:
COMPLETION DATE:

Note. Copyright 2011 by Performance Programs, Inc. Used with permission.

SUMMARY

This chapter has addressed the interplay between individual and organizational assessments, with a particular focus on organizational assessment approaches. Assessment practices communicate care and thoughtfulness and can encourage employee engagement. Individual assessments of cognitive capacity, personality tendency, motive, and behavior can all lead to improved organizational performance. However, organizational practices can also affect individual performance, allowing leaders to create an environment in which individuals can thrive. Organizations that want to have assessment metrics can rely on three types of surveys: audits (to gather information), alignment (to analyze consistency between values and actions), or engagement (to identify and encourage employee involvement in solutions). The six steps of the survey process are a guideline on how to actually implement an organizational survey. Surveys provide information and a structure for communications and forward movement. They offer great potential to increase performance, yet they are neither simple nor magical in producing results. As with every worthwhile endeavor, if executed with care and respect for people, the benefits are a data-rich asset for the organization.

References

Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly, 51*, 220–232. doi:10.1086/269030

Church, A., Waclawski, J., & Kraut, A. (2001). *Designing and using organizational surveys: A seven-step process.* San Francisco, CA: Jossey-Bass.

Cochran, W. G. (1963). *Sampling techniques* (2nd ed.). New York, NY: Wiley.

Connolly, P., & Connolly, K. (2005). *Employee opinion questionnaires: 20 ready-to-use surveys that work.* San Francisco, CA: Pfeiffer.

Connolly, P., & Connolly, K. (2006). *Employee surveys—Practical and proven methods, samples, examples* (2nd ed.). Old Saybrook, CT: Performance Programs.

Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. *Harvard Business Review, 88*, 52–58, 150.

Dunham, R. B., & Smith, F. J. (1979). *Organizational surveys: An internal assessment of organizational health.* Glenview, IL: Scott, Foresman.

Edwards, J. E., Thomas, M. D., Rosenfeld, P., & Booth-Kewley, S. (1996). *How to conduct organizational surveys.* Thousand Oaks, CA: Sage.

Gallup Organization. (1998). Employee research: From nice to know to need to know. *Personnel Journal, 67*, 42–43.

Herzberg, F. (2003). One more time: How do you motivate employees? 1968. *Harvard Business Review, 81*, 87–96.

Higgs, A. C., & Ashworth, S. D. (1996). Organizational surveys: Tools and assessment for research. In A. Kraut (Ed.), *Organizational surveys: Tools for assessment and change* (pp. 19–40). San Francisco, CA: Jossey-Bass.

Hinrichs, J. R. (1996). Feedback, action planning, and follow-through. In A. Kraut (Ed.), *Organizational surveys: Tools for assessment and change* (pp. 255–279). San Francisco, CA: Jossey-Bass.

Hogan, R. T. (2007). *Personality and the fate of organizations.* Mahwah, NJ: Erlbaum.

Kraut, A. (1996). *Organizational surveys: Tools for assessment and change.* San Francisco, CA: Jossey-Bass.

Lancaster, L., & Stillman, D. (2002). *When generations collide: Who they are, why they clash.* New York, NY: HarperCollins.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97. doi:10.1037/h0043158

Ong, A. D., & Weiss, W. J. (2000). The impact of anonymity on responses to sensitive questions. *Journal of Applied Social Psychology, 30,* 1691–1708. doi:10.1111/j.1559-1816.2000.tb02462.x

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65–88. doi:10.1146/annurev.soc.29.110702.110112

Schiemann, W. (1996). Driving change through surveys: Aligning employees, customers, and other key stakeholders. In A. Kraut (Ed.), *Organizational surveys: Tools for assessment and change* (pp. 88–116). San Francisco, CA: Jossey-Bass.

Schneider, B., Salvaggio, A. N., & Subirats, M. (2002). Climate strength: A new direction for climate research. *Journal of Applied Psychology, 87*, 220–229. doi:10.1037/0021-9010.87.2.220

Sudman, S., & Bradburn, N. (1974). *Response effects in surveys: A review and synthesis.* Chicago, IL: Aldine.

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco, CA: Jossey-Bass.

Wagner, D. A., & Spencer, J. J. (1996). The role of surveys in transforming culture. In A. Kraut (Ed.), *Organizational surveys: Tools for assessment and change* (pp. 67–87). San Francisco, CA: Jossey-Bass.

Wiley, J. (2010). *Strategic employee surveys: Evidence-based guidelines for success.* San Francisco, CA: Jossey-Bass.

Witt, L. A., Andrews, M. C., & Kacmar, K. M. (2000). The role of participation in decision-making in the organizational politics–job satisfaction relationship. *Human Relations*, *53,* 341–358. doi:10.1177/0018726700533003

CHAPTER 35

COUNTERPRODUCTIVE WORK BEHAVIORS: CONCEPTS, MEASUREMENT, AND NOMOLOGICAL NETWORK

Deniz S. Ones and Stephan Dilchert

Every day, some individuals engage in behaviors that, rather than adding value to their organization, detract from it. Reasons for committing such acts, ranging from theft to abusing sick leave to violence at work, are diverse and can be situational or dispositional. However, what these behaviors have in common is that they are contrary to the legitimate goals of organizations (Sackett & DeVore, 2001). Empirical evidence has suggested that these individual behaviors are all part of a larger phenomenon of counterproductivity that incurs enormous costs to organizations, societies, and economies worldwide.

In 2003, the U.S. Bureau of Labor Statistics estimated that U.S. companies lost 2.8 million productive work days each year owing to absenteeism. Associated costs were estimated at more than $74 billion. Another U.S. Bureau of Labor Statistics (2006) survey revealed that nearly 5% of U.S. corporations reported incidents of violence in 2005 (incident rates were much higher in government) and that more than a third of those incidents had a negative impact on the workforce. In the private sector, the occurrence rates of violence involving customers and coworkers were about equal, and they increased with organizational size (among organizations with more than 1,000 employees, more than half reported incidents of workplace violence).

The work psychology literature has collectively referred to undesirable employee behaviors as *counterproductive work behavior* (CWB). In the quest to better understand this construct, this chapter offers a measurement-based, quantitative, psychological perspective. The purpose is to provide an overview of CWB and its measurement in work psychology. To this end, first the construct space for CWB is defined, locating it in models of job performance. Second, competing definitions from the literature and the strengths and weaknesses of each are summarized. Third, both broad measures and specific indicators of the construct and its lower order factor structure are described. Fourth, the reliability of scale scores in this domain is reviewed. Fifth, the measurement of CWB using observer reports is discussed. Sixth, findings on individual-differences correlates of CWB measures from the meta-analytic literature are summarized. The authors also offer some recommendations for better conceptualization and measurement of the CWB construct domain.[1]

DEFINING THE CONSTRUCT SPACE

The economic productivity and viability of organizations depend on the productivity of their employees. Economically relevant employee behaviors in work

Both authors contributed equally to this chapter.

[1]Throughout this chapter, in reviewing the relevant literature and drawing empirically based conclusions, we relied almost exclusively on meta-analytic research. Two major reasons drove this decision. First, the literature on CWB is voluminous, especially when one considers the various alternate manifestations of the construct space (e.g., workplace deviance, employee aggression, unethical behavior). Thus, relying on meta-analytic summaries was a practical approach. Second, and more important, by pooling results across multiple studies, meta-analysis produces results that are less affected by sampling error. We also wanted to ensure that the review of the literature was not distorted by other statistical artifacts such as unreliability in measures and various forms of range restriction, among others (see Hunter & Schmidt, 2004).

DOI: 10.1037/14047-035
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

settings are studied under the rubric of job performance. Although the definitions of job performance are numerous (see Austin & Villanova, 1992, for a review), for the purposes of this chapter, the one offered by Viswesvaran and Ones (2000) is a distillation that is especially relevant to our understanding of the CWB construct. Viswesvaran and Ones defined *job performance* as "scalable actions, behavior, and outcomes that employees engage in or bring about that are linked with and contribute to organizational goals" (p. 216). Accordingly, both behaviors that positively contribute to organizational goals and behaviors that detract from achieving organizational goals are included in this definition. The latter constitute counterproductive behaviors.

Campbell's model of performance (see Chapter 22, this volume) also includes both positive (enhancing) and negative (detracting) behaviors. On the basis of extensive research across a variety of jobs in the military, Campbell, Gasser, and Oswald (1996) identified several behavioral clusters making up job performance, among them maintaining personal discipline. This dimension corresponds to the avoidance of counterproductive behaviors and is described as "the degree to which negative behaviors, such as alcohol and substance abuse at work, law and rule infractions, and excessive absenteeism, are avoided" (p. 266).

Three key features can be identified for defining and measuring job performance constructs generally and in terms of their specific indicators: (a) They describe what employees do in work settings, (b) they can be scaled in terms of each employee's net contribution to achieving organizational goals, and (c) they ought to be under the control of the individual performing them (see Campbell et al., 1996, for a detailed description of these features). In the past decade, the move has been toward acknowledging three primary performance domains: task performance, organizational citizenship behavior (OCB), and avoidance of CWB (Sackett, 2002; Viswesvaran & Ones, 2000).[2] *Task performance* includes behavior that is required of employees (Borman & Motowidlo, 1997). OCB (also called *contextual performance* and *prosocial behavior*) refers to discretionary behaviors in work settings (Organ, 1988). Postulated facets of OCB include altruism, courtesy, civic virtue, rule compliance, sportsmanship, and interpersonal versus organizational OCB (see Hoffman & Dilchert, 2012).

Engaging in (and conversely, avoiding) CWB constitutes the third primary component of the job performance construct. This behavior includes a broad variety of phenomena. Employee absenteeism, abusive supervision, aggression, blackmail, bribery, bullying, destruction of property, discrimination, drug use, extortion, fraud, harassment, industrial espionage, interpersonal violence, kickbacks, lying, sabotage, sexual harassment, social loafing at work, social undermining, tardiness, theft, tyranny, violations of confidentiality, violence, and withdrawal behaviors have all been the subject of applied psychological research and are included in the CWB construct space. As Ones and Viswesvaran (2003) noted,

> Such behaviors can be termed (1) counterproductive, as they detract from the productive behaviors at work, (2) disruptive, as they disrupt work-related activities, (3) antisocial, as they violate social norms, and (4) deviant, as they diverge from organizationally desired behaviors. (p. 211)

Although large-scale meta-analytic evidence has supported the existence of a general factor of job performance, some of the variability in job performance can better be modeled by taking the three primary lower level factors (task performance, OCB, and CWB) into consideration. That is, although task performance, OCB, and avoiding CWB are all positively and substantially related to one another, the magnitudes of interrelationships leave room for divergent validity. The overlap between OCB and CWB has been of particular interest; recent meta-analytic work, however, has presented divergent validity evidence for measures of both domains. Cumulating data from 49 studies including 16,721 individuals, Dalal (2005) estimated the true score correlation (i.e., correlation corrected for unreliability in both measures) between OCB and CWB as −.32 (the sample size weighted mean observed correlation was −.27). Subsequent meta-analyses have estimated relationships in the −.20 to −.50 range, depending

[2]Nonetheless, job performance is a hierarchically organized construct with a general factor of job performance at its apex (even after controlling for the effects of halo error, the general factor accounts for 60% of the reliable variance in job performance ratings; Viswesvaran, Schmidt, & Ones, 2005).

on the dimensions of OCB and CWB studied (Berry, Ones, & Sackett, 2007). Of equal import, nomological networks of the two constructs have shown marked differences (see Berry et al., 2007; Dalal, 2005; Hoffman & Dilchert, 2012; Ilies, Fulmer, Spitzmuller, & Johnson, 2009), substantially weakening the argument that OCB and CWB are opposite poles of the same performance dimension.[3]

DEFINING COUNTERPRODUCTIVE WORK BEHAVIOR

Several popular and widely quoted definitions of CWB exist. Robinson and Bennett (1995) defined CWB as "voluntary behavior that violates significant organizational norms and in so doing threatens the well-being of an organization, its members, or both" (p. 556). Similarly, Sackett and DeVore (2002) defined CWB as "intentional behavior on the part of an organizational member viewed by the organization as contrary to its legitimate interests" (p. 145). Spector and Fox (2005) defined CWB as "volitional acts that harm or are intended to harm organizations or people in organizations" (p. 151). These three definitions were careful to include only volitional behaviors and to exclude outcomes. However, the major difference is whether harm is intended.

Intent appears to be at the heart of Spector and Fox's (2005) definition. An example they provided highlights this point well:

> To qualify as CWB, the employee must purposely avoid using safe equipment or procedures, thus behaving in a reckless manner that results in injury, even though the injury itself was not desired. Alternately, the individual might engage in the behavior for the specific purpose of causing harm—for example, to damage equipment. (p. 152)

Thus, according to Spector and Fox, CWB involves behaviors that result from employee choices and that is willful (e.g., retaliatory actions). Although Sackett and DeVore's (2001) definition of CWB also made reference to intentional behaviors, their threshold for inclusion is lower, requiring simply nonaccidental behaviors. Hoffman and Dilchert (2012) argued for carefully broadening these popular CWB definitions. The authors agree with Hoffman and Dilchert's assessment that not only volitional (i.e., nonaccidental) acts can harm organizations and would add that intentional behavior is not the same as behavior and outcomes under the employee's control. Regardless of cognition (i.e., whether explicit reasoning is involved) or motive (e.g., malevolence, clumsiness, habit), behaviors that detract from organizational goals and well-being are counterproductive and thus ought to be considered CWB. One does not have to specify intentional behaviors to exclude accidental outcomes from definitions of CWB.

The following definition that explicitly recognizes CWB as a dimension of job performance can be offered: Counterproductive work behaviors are scalable actions and behaviors that employees engage in that detract from organizational goals or well-being and include behaviors that bring about undesirable consequences for the organization or its stakeholders. In this definition, no reference to organizational norms, intention to harm, or even cognitively reasoned actions is made. Harmful acts committed thoughtlessly, impulsively, and even out of (bad) habit are included among CWB, even when no premeditation occurs. This definition explicitly acknowledges a vast series of empirical findings on the failure to control wayward impulses (Gough, 1971). Furthermore, this definition is in line with recent findings from the unethical behavior literature that suggest better prediction of unethical behavior than of unethical intentions, leading

[3]Also, OCB and CWB are distinct from task performance both conceptually and empirically. Measures of the latter can include both work samples (typically assessing can-do aspects of performance and thus maximal performance) and ratings (traditionally assessing typical performance on job tasks). However, both OCB and CWB assessments focus almost exclusively on typical performance (i.e., how employees behave on a day-to-day basis). Thus, when relationships among the three primary components are examined, links with task performance may be affected by whether the task performance measures in question include maximal performance indicators. When all three constructs are assessed using typical performance measures, intercorrelations among them tend to be stronger (e.g., observed correlations in the .45–.65 range, true score correlations greater than .70; Hoffman, Blair, Meriac, & Woehr, 2007) than when task performance includes maximal performance measures (e.g., Sackett, 2002, estimated the observed relationship between task performance and CWB to be −.19 based on Project A data, in which task performance was measured using work samples and job knowledge tests).

scholars to highlight "a need to more strongly consider a new 'ethical impulse' perspective in addition to the traditional 'ethical calculus perspective'" (Kish-Gephart, Harrison, & Treviño, 2010, p. 1).

MEASUREMENT OF COUNTERPRODUCTIVE WORK BEHAVIOR

Broad Measures of Counterproductive Work Behavior

Most early work on CWB in industrial psychology was focused on solving or alleviating applied problems, such as workplace accidents (e.g., Henig, 1927), employee absences (e.g., Kornhauser & Sharp, 1932), or tardiness (Motley, 1926). Hence, early CWB measurement often relied on individual observations, single-item measures, or information coded from personnel records (Hoffman & Dilchert, 2012), which emphasized easily quantifiable outcomes rather than broad measures that assessed a comprehensive spectrum of counterproductivity, which in turn made it difficult to identify generalizable patterns and build comprehensive theories of counterproductivity (Robinson & Bennett, 1995). However, as scholars shifted their focus from predicting to understanding counterproductive behaviors at work, a variety of broader, multi-item, and multidimensional measures were developed. Most measures included a list of several undesirable behaviors for which employees had to report either a frequency or whether they had ever engaged in them.

One of the earliest examples of a comprehensive CWB scale in the work psychology literature is that of Spector (1975). Using employee self-report data, Spector identified six interpretable factors among 35 CWB items: aggression against others, sabotage, wasting of time and materials, interpersonal hostility and complaining, interpersonal aggression, and apathy about the job. Other scales followed that operationalized CWB as a broad phenomenon of qualitatively distinct but empirically related behaviors (e.g., Robinson & Bennett, 1995; see the next section). Throughout the 1990s, CWB measures proliferated, with a variety of construct labels and measures with little cross-pollination. Table 35.1 presents a selection of counterproductivity variables that have been proposed. For most of these constructs and corresponding scales, no divergent validity evidence has shown them to be distinct from other broad measures of CWB.

Table 35.1 also illustrates a more recent trend in CWB research. It lists a variety of narrower constructs that have been coined lately. In contrast to early CWB measures that focused on single observations of concrete CWB, these variables often involve multi-item measures akin to broad CWB scales discussed earlier, yet are limited to a narrow behavioral phenomenon (e.g., mobbing). Again, the problem is a proliferation of labels in parallel literatures that often do not build on prior knowledge of the unifying CWB construct. In those literatures, few attempts have been made to provide divergent or convergent validity evidence with other aspects of CWB. Although the development and refinement of more nuanced scales can potentially make a contribution to CWB research, the troublesome aspect of this development is the failure of many scholars to adequately position their scales in the wider nomological network and provide empirical evidence of their constructs' validity.

Lower Order Structure of the Counterproductive Work Behavior Domain

One taxonomic and scale development effort that has received a considerable amount of attention is that of Robinson and Bennett (1995), who used multidimensional scaling to establish their typology of deviant workplace behaviors. Robinson and Bennett first gathered 45 critical incidents of deviant behavior at work observed by a sample of office personnel and employed MBA students. A second sample judged a subset of critical incident pairs with regard to their perceived similarity. Robinson and Bennett's interpretation of their multidimensional scaling results suggested that a two-dimensional model fit these data better than a one-dimensional one and that models of a higher order provided no notable improvement; Robinson and Bennett thus pursued the more parsimonious, two-dimensional model. They proceeded to examine subjects' descriptions of the most common attributes that described the 45 deviant behaviors and used judges' ratings of these attributes to identify the meaning of

TABLE 35.1

Examples of Broad and Narrow Counterproductive Work Behavior (CWB) Conceptualizations

Variable	Description or definition	Source
	Broad CWB conceptualizations	
Antisocial behavior	Any behavior that brings harm or is intended to bring harm to the organization, its employees, or its stakeholders	Giacalone & Greenberg (1997)
Employee vice	An act that betrays the trust of either individuals or the organizational community	Moberg (1997)
General counterproductive behavior	Behaviors that violate the legitimate interests of an organization by being potentially harmful	Marcus & Schuler (2004)
Noncompliant behavior	Nontask behaviors that have negative organizational implications	Puffer (1987)
Organization-motivated aggression	Attempted injurious or destructive behavior	O'Leary-Kelly, Griffin, & Glew (1996)
Organizational misbehavior	Any action that violates core organizational or societal norms	Vardi & Wiener (1996)
Organizational retaliation behaviors	Adverse reactions to perceived unfairness	Skarlicki & Folger (1997)
Unethical behavior	Any organizational member action that violates widely accepted (societal) norms	Kish-Gephart, Harrison, & Treviño (2010)
Workplace aggression	Any form of behavior that is intended to harm current or previous coworkers or their organization	Baron & Neuman (1996)
Workplace deviance	Behavior that violates significant organizational norms and, in so doing, threatens the well-being of the organization or its members	Robinson & Bennett (1995)
	Recent narrower CWB conceptualizations	
Abusive supervision	The extent to which supervisors engage in sustained display of hostile verbal and nonverbal behaviors, excluding physical contact	Tepper (2000)
Destructive leadership behavior	Systematic and repeated behavior by a leader, supervisor, or manager that violates the legitimate interest of the organization by undermining or sabotaging the organization's goals, tasks, resources, and effectiveness or the motivation, well-being, or job satisfaction of his or her subordinates	Einarsen, Aasland, & Skogstad (2007)
Mobbing or bullying	Social interaction through which one individual (seldom more) is attacked by one or more (seldom more than four) individuals almost on a daily basis and for periods of many months, bringing the person into an almost helpless position with potentially high risk of expulsion	Leymann (1996); Zapf, Einarsen, Hoel, & Vartia (2003)
Social undermining	Behavior intended to hinder, over time, the ability to establish and maintain positive interpersonal relationships, work-related success, and favorable reputation	Duffy, Ganster, & Pagon (2002)
Time banditry	Propensity of employees to engage in non-work related activities during work time	Martin, Brock, Buckley, & Ketchen (2010)
Workplace incivility	Low-intensity deviant behavior with ambiguous intent to harm the target, in violation of workplace norms for mutual respect	Andersson & Pearson (1999)

Note. The information on broad measures of CWB was adapted from Ones, Connelly, Viswesvaran, and Salgado (2008).

the two underlying dimensions. The first replicated earlier reports (Wheeler, 1976) that CWB can be distinguished on a continuum ranging from minor to serious. The second, however, distinguished between deviant acts targeted at the organization versus interpersonally (e.g., against coworkers or supervisors). They termed the four quadrants *property deviance, production deviance, personal aggression,* and *political deviance.*

Perhaps the most important contribution of Robinson and Bennett's (1995) model is the distinction between interpersonally and organizationally targeted counterproductive work behaviors (here, CWB–I and CWB–O, respectively), which has received wide attention in the research literature. Bennett and Robinson (2000) provided a measurement scale that reflects this distinction, which is now extensively use in CWB research. Its popularity is driven, at least in part, by the fact that the CWB–I/CWB–O distinction is intuitively appealing to many and by the fact that the measure is easily available.

Two main issues have been raised regarding the viability and usefulness of this two-dimensional model of CWB. First, Dalal's (2005) meta-analysis established a strong overlap between interpersonal and organizational CWB ($\rho = .70$, $k = 20$, $N = 4{,}136$). Berry et al.'s (2007) meta-analysis reported a similarly strong construct-level relationship ($\rho = .62$, $k = 27$, $N = 10{,}104$), even though the two aspects of CWB have somewhat differing dispositional and situational antecedents. Second, the development of the Robinson and Bennett (1995) taxonomy was based only on perceived similarity between different deviant behaviors, not on empirical evidence of their co-occurrence. This problem is a vexing but common one for taxonomic efforts in the CWB domain (Gruys & Sackett, 2003). The main reason is a concern regarding impression management (i.e., underreporting) when surveying individuals about the counterproductive behaviors they actually engage in at work. However, more recent studies using different sources of CWB measurement (other ratings, objective records) have yielded results that replicate the distinction between CWB–I and CWB–O. For instance, Stewart, Bing, Davison, Woehr, and McIntyre (2009) used observer reports of CWB (using Bennett & Robinson's, 2000, items); one of the three factors describing their data distinguished interpersonally from organizationally targeted deviance. Dilchert, Ones, Davis, and Rostow (2007) also replicated the CWB–I/CWB–O distinction and did so using objectively reported data on co-occurrence of many qualitatively different CWB (e.g., sexual harassment, at-fault motor vehicle accidents, violence). They relied on detected incidents of employee counterproductivity (reflected in the personnel files of 1,700 police officers). The correlation between the CWB–I and CWB–O factors was .52 (uncorrected), which corresponds precisely to the two meta-analytic estimates for self-report data (Berry et al., 2007; Dalal, 2005). Although it is clear that the two dimensions are far from distinct and that both correlate highly with the overarching factor, distinguishing between the target of CWB provides utility in current theoretical models and also seems to have applied use given differential patterns of dispositional and situational antecedents (see Berry et al., 2007).

Reliability of Counterproductive Work Behavior Measurement

Ones, Viswesvaran, and Schmidt (1993) reported a mean sample-size-weighted internal consistency reliability of .69 for scores on CWB measures. Dalal (2005) pooled internal consistency reliabilities across 49 studies ($N = 16{,}721$) and found that the sample-size-weighted mean reliability coefficient for overall CWB measures was .77. The unknown composition of CWB included in the respective studies pooled makes direct comparisons difficult. Furthermore, the types of reliability estimates included in these analyses likely varied (e.g., alpha, composite alpha, Mosier). However, taking the higher value of .77, it is clear that even perfect relationships with CWB (construct level correlations of 1.00) would be attenuated to a maximal value of .88 owing to unreliability in CWB measurement ($\sqrt{.77} = .88$). For interpersonally targeted CWB measures, sample size weighted reliability estimates have been reported as .68 ($k = 20$, $N = 4{,}136$; Dalal, 2005) and .84 ($k = 27$, $N = 6{,}357$; Berry et al., 2007). For CWB targeting organizations, the sample size weighted reliability estimates have

been reported as .77 ($k = 27$, $N = 6{,}357$; Dalal, 2005) and .82 ($k = 22$, $N = 6{,}080$; Berry et al., 2007).[4]

Interestingly, most of the CWB reliabilities included in meta-analytic estimates have been computed for self-report measures. Viswesvaran, Ones, and Schmidt (1996) reported a mean internal consistency reliability of .77 ($k = 15$, $N = 3{,}438$) for supervisory ratings of CWB. The mean interrater reliability estimate of CWB ratings has been reported as .58 for supervisors and .71 for peers (Viswesvaran et al., 1996). Pooling across the 13 studies contributing to these estimates ($N = 1{,}125$), the overall interrater reliability is .67 (reliability of a single rater). Future research should assess the temporal stability of scores on counterproductive behavior measures in work settings. Stability versus dynamicity of CWB's relationships with other constructs is a function of their temporal trajectory. Although Dalal, Lam, Weiss, Welch, and Hulin (2009) examined dynamic relationships of work behaviors, including CWB, their approach was geared to understanding within-person variability in the variables examined. Studying CWB in association with other antisocial behaviors across the lifespan can provide unique insights and opportunities to work psychologists

Measurement Using Observer Reports of Counterproductive Work Behavior

In applied settings, CWB measurement (especially that of narrow facets such as absenteeism, employee theft, etc.) has traditionally relied on organizational records, both because of the relative ease of measurement and the fact that such information can easily be linked to other quantitative indicators of individual and organizational productivity. Even though such data are very meaningful to organizations, to the degree that they capture outcomes of counterproductive behaviors they may only be distally related to employee behavior (see the section Defining Counterproductive Work Behavior earlier in this chapter). From the point of CWB research and theory development, such measures (especially those that capture only single facets of CWB) also raise concerns about criterion deficiency and undesirable measurement properties resulting from the problem of low base rates. With the advent of multi-item, broad measures of CWB, the measurement mode, at least for research purposes, has shifted such that the vast majority of studies now use participant self-reports.

Hoffman and Dilchert (2012) discussed the important role that self-reports play in CWB research. First, they circumvent the problem of low base rates often observed in objective or outcome measures of CWB. Ultimately, the most knowledgeable source for information on employees' on-the-job behaviors are the employees themselves (Fox, Spector, Goh, & Bruursema, 2007). Hence, under conditions of confidentiality (and notwithstanding effects such as self-denial or memory loss), an employee should be able to provide the most accurate criterion measurement with regard to the frequency of CWBs engaged in over time. Because organizational records often suffer from low base rates because of undetected incidents and supervisor ratings suffer from a lack of opportunity to observe, self-reports play an important role when measuring CWB for research purposes.

However, as with any criterion measure, it is important to consider the accuracy of self-report CWB measurement. Given that CWB constitutes behaviors that result in harm, underreporting on behalf of employees is the prime concern. For most research applications, however, the question is not whether potential underreporting will affect the reported base rates of CWB, but whether such response behavior might distort the rank order of employees on the CWB measure (particularly relevant for validation purposes). Hence, recent efforts to expand CWB measurement using other reports, in particular, supervisor and peer ratings (see de Jonge & Peeters, 2009; Fox et al., 2007; Stewart et al., 2009) might add value to broad CWB measurement. In this regard, it is important to evaluate the overlap of self-reports with alternate measurement sources.

[4]The discrepancy between Dalal's (2005) and Berry et al.'s (2007) findings are the result of the specific measures included in the respective summaries. Berry et al. included only those measures in their analyses that were broad measures of CWB that directly targeted the measurement of interpersonal and organizational CWB (e.g., Bennett & Robinson, 2000). Dalal included measures of varying specificity that could be classified as either interpersonal CWB or organizational CWB.

Knowledge of self–other rating overlap in CWB measures addresses two important questions: First, when employees report on their own CWBs (even under conditions of confidentiality), do they provide information that concurs with that of other sources, strengthening the confidence one has in such potentially motivated self-disclosures? (Conversely, one might ask whether, under conditions of confidentiality, the overlap of self- and other reports is strong enough so that one might substitute self-reports with observer ratings when confidentiality cannot be guaranteed.) Second, do other reports potentially add informational value over and above that already contained in self-reports? That is, do other or multisource ratings pose the potential to increase the quality of the research on and knowledge of counterproductivity at work? As Hoffman and Dilchert (2012) pointed out, the relative ease of collecting self-reports of CWB, and the strong theoretical arguments for the quality of employee-reported information (knowledge of actual behavior rather than opportunity to observe outcomes), mean that currently no substitute for self-report measures in CWB research exists. However, an external perspective might possibly supplement existing types of CWB measures in a meaningful way.

Berry, Carpenter, and Barrett (2010) recently provided a meta-analysis that sheds light on this question and leads to some provocative conclusions. Berry et al. cumulated data on the overlap between self-reports and CWB ratings from different sources (peers and supervisors, collectively termed *other ratings*). They showed that self-reports on other ratings of CWB were moderately to strongly related. More important, data from the two types of sources exhibited similar correlational patterns with a variety of external variables. One notable finding, however, was that self- and other reports correlated more highly for interpersonally targeted CWB than organizationally targeted CWB. The authors have previously argued that the opportunity to observe CWB varies across CWB domains. By definition, interpersonal CWB is harder to conceal because "typically, there is at least one person other than the perpetrator involved (e.g., the victims of sexual harassment, violence, racially offensive conduct and behavior . . .)" (Dilchert et al., 2007, p. 625). Nonetheless, the mostly consistent pattern of external correlates of self-reports and other ratings of CWB from Berry et al.'s analysis is intriguing and poses questions of the utility of collecting other reports. They also point out that self-reports capture a wider spectrum of CWB (i.e., others, particularly supervisors, report observing fewer CWBs than employees report themselves). Their meta-analysis revealed that self-reported mean levels of CWB were higher on average than those reported by observers. Berry et al. also showed that other ratings account for small amounts of incremental variance in external correlates above self-reports of CWB and hence conclude that the administrative difficulties associated with collecting such ratings might often not be justified. For research applications, especially when participants' confidentiality can be guaranteed and employees are convinced that they can respond honestly without fear of negative consequences, self-reports currently offer an acceptable approach to CWB measurement.

UNDERSTANDING THE NOMOLOGICAL NETWORK OF COUNTERPRODUCTIVE WORK BEHAVIORS MEASURES

Correlates and determinants of CWB are found in both individual differences characteristics and contextual factors. The former constitute the potential psychological capital and liabilities that individuals bring to their work environments. The latter capture the environmental influences that contribute to CWB. In this section, we describe the nomological network in terms of person-based correlates to enrich understanding of counterproductivity at work. Because of space constraints, the interested reader is referred to Spector and Rotundo (2010) for a summary of organizational justice variables, stressors, and their influence on CWB. Here, readers first find a brief overview of relations with psychological individual differences variables, followed by a review of demographic variables and CWB.

Relationships With Psychologically Based Individual Differences

The relationships of CWB to personality characteristics, cognitive ability, and individual cultural values

have been systematically studied and reported since the early 1990s. Recent qualitative and quantitative reviews of the literature have already provided much detail on CWB's relationship with personality (Berry et al., 2007; Hoffman & Dilchert, 2012; Ones & Viswesvaran, 2003, 2011; Rotundo & Spector, 2010). Thus, only a very brief overview is provided here, and the reader is directed to these sources for more detailed discussions.

Of the Big Five dimensions of personality, conscientiousness is perhaps the one most closely linked with avoidance of CWB. Both Salgado's (2002) and Dalal's (2005) meta-analyses estimated the conscientiousness–CWB true score relationship. Salgado's estimate was $\rho = -.26$ ($k = 13$, $N = 6276$, $r = -.16$), whereas Dalal's estimate was $\rho = -.38$ ($k = 10$, $N = 3{,}280$, $r = -.29$). Berry et al. (2007) conceptualized CWB as explicitly incorporating interpersonal and organizational counterproductivity and estimated the relationship with conscientiousness as a ρ of $-.35$. However, the relationship with CWB–O was stronger at $\rho = -.42$ ($k = 8$, $N = 2{,}934$, $r = -.34$) than the relationship with CWB–I ($\rho = -.23$, $k = 11$, $N = 3{,}458$, $r = -.19$). Conscientiousness is more closely related to CWB–O than to CWB–I. Furthermore, a meta-analysis focusing on conscientiousness facets has reported that the dependability facet of the trait is more closely related to CWB than are the achievement, order, and cautiousness facets (Dudley, Orvis, Lebiecki, & Cortina, 2006). On the basis of another two separate meta-analyses, conscientiousness also appears to be moderately related to at least one of the specific counterproductivity outcomes, accidents (Christian, Bradley, Wallace, & Burke, 2009, corrected correlation $-.26$; Clarke & Robertson, 2005, corrected correlation $= -.27$).

Agreeableness has also been related to avoidance of CWB. Salgado (2002) estimated the agreeableness–CWB true score level relationship to be $-.20$ ($k = 9$, $N = 1{,}299$, $r = .13$). Berry et al.'s (2007) estimate using the CWB–I/CWB–O composite was higher at $-.44$. Agreeableness was a better predictor of CWB–I than CWB–O. The respective true score correlations were estimated to be $-.46$ ($k = 10$, $N = 3{,}336$, $r = -.36$) and $-.32$ ($k = 8$, $N = 2{,}934$, $r = -.25$). An interesting question is whether different forms of interpersonal CWB might be differentially predicted by agreeableness.

The personality trait of neuroticism is also related to CWB, although at lower levels than conscientiousness and agreeableness. The meta-analytic estimate of the true score correlation is .26 (Berry et al., 2007). In contrast to conscientiousness and agreeableness, neuroticism relates similarly to CWB–I and CWB–O (.24 and .23, respectively). The construct of negative affect, which is related to neuroticism, also appears to be related to CWB. Three meta-analytic (unreliability corrected) estimates have been reported: .41 (Dalal, 2005), .29 and .28 (CWB–I and CWB–O, respectively; Herchovis et al., 2007), and .30 (Kaplan, Bradley, Luchman, & Haynes, 2009). Thus, the magnitude of the relationships of negative affect with CWB appears to be somewhat larger than those typically found for neuroticism.

Relationships of CWB with extraversion and openness to experience have either been negligible (Berry et al., 2007) or extremely variable (Salgado, 2002). However, the positive affect facet of extraversion has been reported to correlate $-.34$ (corrected for unreliability) with CWB (Dalal, 2005). Future research can benefit from examining whether the sensation-seeking facet of extraversion, which is conceptually related to several aspects of CWB, also displays useful relationships.

In addition to the Big Five dimensions and facet-level personality constructs just described, compound personality traits have been shown to relate strongly to counterproductivity at work (Ones, Viswesvaran, & Dilchert, 2005). Such compound personality measures include integrity tests (Ones, Viswesvaran, & Schmidt, 2003), customer service scales (Ones & Viswesvaran, 2008), violence scales (Ones & Viswesvaran, 2001b), drug and alcohol scales (Ones & Viswesvaran, 2001a), and stress tolerance scales (Ones & Viswesvaran, 2011). The relationships of these scales with CWB and its facets tend to be among the strongest for individual differences traits. All these measures assess compound personality traits defined by conscientiousness, agreeableness, and emotional stability to varying degrees (Ones, 1993). Given the literature reviewed earlier, which showed that

conscientiousness, agreeableness, and emotional stability appear to function as antecedents of CWB, it is not surprising that personality measures constructed specifically to predict CWB and relevant behaviors also tap into these three constructs. Additional compound traits for which meta-analyses have reported relationships include locus of control and Machiavellianism (unreliability corrected correlation of .25 with unethical behaviors in both cases; Kish-Gephart et al., 2010).

Until recently, the relationship between cognitive ability and CWB had been relatively unexamined. Few, inconsistent studies in this area either measured cognitive ability earlier in life (Roberts, Harms, Caspi, & Moffitt, 2007) or used small and restricted samples (Marcus & Schuler, 2004). However, three large-scale studies examining cognitive ability's relationships with CWB have found moderate relationships on par with those reported for conscientiousness and compound personality scales (Dilchert et al., 2007; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Oppler, McCloy, & Campbell, 2001). Cognitive ability may have an inhibitory effect that keeps individuals from engaging in CWB.[5] Future models of CWB should include cognitive ability as one of the determinants of CWB.

Finally, individual-level cultural values have also recently been postulated to explain variance in CWB (Taras, Kirkman, & Steel, 2010). Higher scores on individualism and uncertainty avoidance and lower scores on power distance and masculinity are related to avoiding unethical behavior. The following relationships were reported in Taras et al.'s comprehensive meta-analysis of cultural values: Individualistic values correlated .39 with avoiding CWB (collectivistic values were associated with higher levels of CWB). Similarly, individuals scoring high on uncertainty avoidance engaged in less CWB (.20). Lower scores on power distance were correlated .38 with CWB. Finally, feminine cultural values correlated .38 with avoiding CWB (all values are corrected for unreliability). Contemporary models of CWB may need to take into account cultural values as well.

Relationships With Demographic Variables

Table 35.2 summarizes the relationships between demographic variables and CWB found in the meta-analytic literature. Demographic variables for which meta-analytic data could be located were age, tenure, work experience, educational level, and gender; relationships with these demographic variables are modest.

Age displays small negative relationships with CWB. That is, older individuals appear to engage in less CWB than younger individuals. Although the difference between CWB–I and CWB–O in the way each relates to age does not appear to be large, a few specific domains of CWB display stronger negative relationships: production deviance ($-.33$, $k = 3$, $N = 9{,}175$) and theft ($-.21$, $k = 3$, $N = 9{,}175$) as well as tardiness. For the latter, depending on the specific meta-analysis and operationalization of tardiness, correlations ranged between $-.12$ and $-.28$. Some of these negative correlations could be the result of older employees having been caught and dismissed from the organization on the basis of CWB.

Relationships with tenure were also mostly small and negative. In meta-analyses that distinguished between self-reported versus externally detected CWB (i.e., supervisor ratings, peer ratings, organizational records; Ng & Feldman, 2010), the negative correlation of tenure was stronger with external records and ratings. Taken at their face value, these results would appear to lead to the conclusion that employees with less tenure engage in somewhat higher levels of CWB. However, when age was controlled for, tenure tended to relate positively to CWB (Ng & Feldman, 2010). This finding may be partially because longer tenured employees have more and different opportunities to engage in CWB (opportunity effect), the cumulative nature of CWB over time (toxic accumulation effect), and the norm violation entitlements afforded to longer tenured employees (blind organizational eye effect).

Work experience displayed a true score correlation of $-.20$ with CWB, although relationships with organizational CWB appear to be somewhat

[5]See Dilchert et al. (2007) for a discussion of the role of status variables such as socioeconomic status and educational level, which are often postulated to cause a spurious relationship between intelligence and delinquency.

TABLE 35.2

Demographic Variables and CWB: Summary of Findings From Meta-Analytic Investigations

CWB criterion	Meta-analytic source	*k*	*N*	*r*	ρ	SD_ρ	80% CI
			Age				
CWB overall ("unethical behavior")	Kish-Gephart, Harrison, & Treviño (2010)	17	5,034	−.10	**−.11**	.09	[−.23, .01]
CWB overall (self-report)	Ng & Feldman (2008)	28	7,072	—	**.12**	.08	[.02, .22]
CWB overall (supervisor & peer rated)	Ng & Feldman (2008)	6	1,151	—	**−.12**	.08	[−.22, −.02]
CWB overall (composite)[b]	Berry, Ones, & Sackett (2007)	—	—	−.08	**−.09**	—	—
CWB–interpersonal	Berry et al. (2007)	14	6,249	−.05	**−.06**	.06	[−.13, .02]
Workplace aggression (self-report)	Ng & Feldman (2008)	15	3,641	—	**−.08**	.15	[−.27, .11]
CWB–organizational	Berry et al. (2007)	12	5,928	−.09	**−.10**	.08	[.21, .00]
Absenteeism	Lau, Au, & Ho (2003)[a]	6	1,221	−.11	**−.11**	.07	[−.20, −.02]
Absenteeism (nonsickness, self-report)	Ng & Feldman (2008)	6	3,024	—	**−.01**	.11	[−.15, .13]
Absenteeism (nonsickness, objective measures)	Ng & Feldman (2008)	12	2,508	—	**−.10**	.20	[−.36, .16]
Noncompliance with safety rules	Ng & Feldman (2008)	5	612	—	**−.10**	.07	[−.19, −.01]
Production deviance	Lau et al. (2003)[a]	3	9,175	−.33	**−.33**	.06	[−.41, −.26]
Substance abuse (on the job, self-report)	Ng & Feldman (2008)	15	5,182	—	**−.07**	.08	[−.17, .03]
Tardiness	Koslowsky, Sagie, Krausz, & Singer (1997)	7	1,713	−.15	**−.19**	.13	[−.35, −.03]
Tardiness	Lau et al. (2003)[a]	3	391	−.20	**−.21**	.07	[−.30, −.12]
Tardiness (self-report)	Ng & Feldman (2008)	7	1,657	—	**−.12**	.12	[−.27, .03]
Tardiness (externally detected)	Ng & Feldman (2008)	7	1,763	—	**−.28**	.14	[−.46, −.10]
Theft	Lau et al. (2003)[a]	3	9,175	−.21	**−.21**	.04	[−.26, −.17]
			Tenure				
CWB overall (self-report)	Ng & Feldman (2010)	19	5,357	—	**−.05**	.04	[−.10, .00]
CWB overall (supervisor rating)	Ng & Feldman (2010)	4	1,478	—	**−.19**	.07	[−.28, −.10]
CWB overall (peer or other rating)	Ng & Feldman (2010)	3	440	—	**.11**	.00	[.11, .11]
CWB overall (organizational records)	Ng & Feldman (2010)	3	326	—	**−.14**	.00	[−.14, −.14]
CWB overall (composite)[b]	Berry et al. (2007)	—	—	−.05	**−.05**	—	—
CWB–interpersonal	Berry et al. (2007)	7	2,211	−.01	**−.01**	.00	[−.01, −.01]
Workplace aggression (self-report)	Ng & Feldman (2010)	7	1,696	—	**.07**	.10	[−.06, .20]
CWB–organizational	Berry et al. (2007)	9	2,710	−.07	**−.08**	.05	[−.14, −.01]
Absenteeism	Lau et al. (2003)[a]	4	1,807	−.05	**−.13**	.12	[−.28, .02]
Absenteeism (nonsickness, organizational records)	Ng & Feldman (2010)	14	56,708	—	**−.18**	.05	[−.24, −.12]

(*Continued*)

TABLE 35.2 (*Continued*)

Demographic Variables and CWB: Summary of Findings From Meta-Analytic Investigations

CWB criterion	Meta-analytic source	*k*	*N*	*r*	ρ	SD_ρ	80% CI
Absenteeism (nonsickness, organizational records)	Ng & Feldman (2010)—outlier removed	13	4,734	—	**−.04**	.10	[−.17, .09]
Substance abuse (on the job, self-report)	Ng & Feldman (2010)	16	8,929	—	**−.01**	.04	[−.06, .04]
Tardiness	Koslowsky et al. (1997)	8	2,542	−.06	**−.08**	.08	[−.18, .02]
Tardiness	Lau et al. (2003)[a]	3	391	−.13	**−.13**	.02	[−.15, −.11]
Tardiness (self-report)	Ng & Feldman (2010)	5	904	—	**−.10**	.10	[−.23, .03]
Tardiness (organizational records)	Ng & Feldman (2010)	12	2,629	—	**−.02**	.13	[−.19, .15]
Theft	Lau et al. (2003)[a]	3	9,175	−.10	**−.12**	.03	[−.16, −.08]
	Work experience						
CWB overall (composite)[b]	Berry et al. (2007)	—	—	−.18	**−.20**	—	—
CWB–interpersonal	Berry et al. (2007)	3	794	−.10	**−.11**	.00	[−.11, −.11]
CWB–organizational	Berry et al. (2007)	3	783	−.22	**−.25**	.00	[−.25, −.25]
	Education						
CWB overall	Kish-Gephart et al. (2010)	7	2,621	.00	**.00**	.04	[−.06, .06]
CWB overall (self-report)	Ng & Feldman (2009)[a]	12	3,529	.01	**.01**	.08	[−.09, .11]
CWB overall (supervisor or peer rating)	Ng & Feldman (2009)[a]	7	4,158	−.02	**−.04**	.18	[−.27, .19]
CWB–interpersonal							
Workplace aggression (self-report)	Ng & Feldman (2009)[a]	9	1,801	−.05	**−.09**	.04	[−.14, −.04]
CWB–organizational							
Absenteeism	Lau et al. (2003)[a]	2	269	−.01	**−.01**	.01	[−.02, .00]
Absenteeism (nonsickness, self-report)	Ng & Feldman (2009)[a]	3	957	.02	**.04**	.06	[−.04, .12]
Absenteeism (nonsickness, objective measures)	Ng & Feldman (2009)[a]	6	1,372	−.03	**−.07**	.04	[−.12, −.02]
Substance abuse (on the job, self-report)	Ng & Feldman (2009)[a]	10	11,515	−.17	**−.28**	.11	[−.42, −.14]
Tardiness	Lau et al. (2003)[a]	2	259	−.02	**−.01**	.02	[−.04, .01]
Tardiness (self-report)	Ng & Feldman (2009)[a]	12	6,117	.02	**.04**	.23	[−.25, .33]
Tardiness (externally detected)	Ng & Feldman (2009)[a]	4	645	.02	**.03**	.15	[−.16, .22]

Gender							
				d	δ	SD_δ	
CWB overall	Kish-Gephart et al. (2010)[c]	17	5,350	.18	**.20**	.22	[−.08, .48]
CWB overall (composite)[b]	Berry et al. (2007)[d]	—	—	.29	**.30**	—	—
CWB–interpersonal	Berry et al. (2007)[d]	14	6,250	.28	**.30**	—	[.14, .47]
Interpersonal aggression	Hershcovis et al. (2007)	14	3,653	.39	**.43**	—	[.28, .61]
CWB–organizational	Berry et al. (2007)[d]	12	5,929	.22	**.24**	—	[.04, .45]
Absenteeism	Lau et al.(2003)[a, c]	6	2,093	−.18	**−.20**	—	[−.37, −.04]
Organizational aggression	Hershcovis et al. (2007)	11	3,363	.22	**.26**	—	[−.18, .72]
Tardiness	Koslowsky et al. (1997)[c]	5	1,502	.06	**.08**	—	—
Tardiness	Lau et al.(2003)[a, c]	2	215	−.14	**−.08**	—	[−.49, .35]

Note. CWB = counterproductive work behavior. N = total sample size; k = number of studies; r = sample size weighted mean correlation; ρ = estimate of true-score correlation (corrected for attenuation due to measurement error in CWB measure); $SD\rho$ = standard deviation of ρ; 80% CI = 80% credibility interval (computed by the authors when not available in the original meta-analytic source); d = Cohen's d, computed on the basis of point-biserial correlations (positive values indicate men scored higher on average); δ = estimate of group mean-score difference corrected for attenuation resulting from measurement error in CWB measure and range restriction or enhancement where applicable.

[a]Also corrected for range restriction or enhancement, when applicable.

[b]Composite correlation computed by the authors on the basis of results for CWB–interpersonal and CWB–organizational and their intercorrelation reported in the original meta-analytic source.

[c]Distribution of two groups for original point-biserial correlations unknown (not corrected for).

[d]Based on point-biserial distribution that was corrected for uneven split.

stronger. Given the findings for tenure reviewed earlier, it may be valuable to disentangle these work experience effects from age effects in future research.

The weakest relationships with CWB among demographic variables are reported for educational level. Most meta-analytic estimates were between −.09 and .04 (see Table 35.2) with the exception of self-reported on-the-job substance abuse −.28 ($k = 10$, $N = 11{,}515$, $r = -.17$), indicating that more educated individuals tend to engage in less substance abuse at work.

Finally, Table 35.2 summarizes gender differences on CWB variables. Rather than presenting correlations, standardized mean-score differences between the sexes on CWB scales (Cohen's *d* and corrected δ) are provided. A caveat in interpreting these effect sizes is that with the exception of Berry et al. (2007), most meta-analyses pooled point-biserial correlations. Such correlations are affected by the proportion of men and women in the respective samples and therefore may underestimate true relationships to the extent that the proportions deviate from a 50–50 split. Because only the Berry et al. meta-analysis corrected for this effect, the estimates of gender differences in Table 35.2 might be conservative. Nonetheless, women appear to engage in CWB at moderately lower rates than men. This effect seems to be somewhat stronger for CWB–I than for CWB–O). Unreliability-corrected *d* values were .43 (Hershcovis et al., 2007) and .30 for CWB–I compared with .24 for CWB–O (Berry et al., 2007).

DISCUSSION AND CONCLUSIONS

In this chapter, the authors presented an overview of the conceptualization, measurement, and nomological network of the CWB construct. CWB is a primary dimension of job performance that is distinct from both task performance and organizational citizenship. Various definitions of CWB were reviewed, and strengths and weaknesses of each were identified. A new definition of CWB is provided, one that embeds the construct more directly into existing models of performance and that is more in line with the accumulated empirical findings. The recent proliferation of "new" CWB constructs leads to unnecessary fragmentation of the literature and hampers the generation of cumulative knowledge. The research literatures on many of these modish constructs grow in isolated silos (favoring idiosyncratic theories and antecedent variables) with few attempts to advance understanding of the CWB domain. The field needs a comprehensive, quantitative investigation of how various CWB constructs (general and specific, old and new) relate to one another in employee samples. Because dozens of CWB constructs have been proposed, defined, and measured, a meta-analytic approach may prove most useful in this regard.

Measurement reliability of CWB scale scores appears to be adequate and on par with the reliability in measuring other job performance constructs. The interrater reliability of CWB measures is higher than that for overall job performance and other performance facets (see Viswesvaran, Schmidt, & Ones, 2005). Using other reports to assess CWB is a viable option and might be useful for counteracting some potential disadvantages of self-reports. However, self-report measures of CWB still play a crucial role in a domain in which the subjects themselves most often have better knowledge of their own behaviors and appear to share them unabashedly, especially under conditions of confidentiality.

CWB is clearly associated with specific individual differences and, to a lesser degree, demographic variables. Compound personality scales that amalgamate constructs from conscientiousness, agreeableness, and emotional stability domains (e.g., integrity tests) are especially helpful in predicting CWB in organizational settings. In addition to personality measures, cognitive ability and individual cultural values may also be valuable in identifying individuals likely to engage in CWB.

Counterproductive behaviors are naturally of great concern to organizations, economies, and societies at large. The burgeoning research literature on the CWB construct and its measures that has emerged in the past 20 years is impressive. The future of CWB research is bright.

References

Andersson, L. M., & Pearson, C. M. (1999). Tit for tat? The spiraling effect of incivility in the workplace. *Academy of Management Review, 24*, 452–471.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77*, 836–874. doi:10.1037/0021-9010.77.6.836

Baron, R. A., & Neuman, J. H. (1996). Workplace violence and workplace aggression: Evidence on their relative frequency and potential causes. *Aggressive Behavior, 22*, 161–173. doi:10.1002/(SICI)1098-2337(1996)22:3<161::AID-AB1>3.0.CO;2-Q

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*, 349–360. doi:10.1037/0021-9010.85.3.349

Berry, C. M., Carpenter, N., & Barrett, C. (2010, May). *Self-reports vs. non-self-reports of counterproductive work behavior: A meta-analysis*. Poster presented at the annual conference of the Association for Psychological Science, Boston, MA.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92*, 410–424. doi:10.1037/0021-9010.92.2.410

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99–109. doi:10.1207/s15327043hup1002_3

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 258–299). San Francisco, CA: Jossey-Bass.

Christian, M. S., Bradley, J. C., Wallace, J. C., & Burke, M. J. (2009). Workplace safety: A meta-analysis of the roles of person and situation factors. *Journal of Applied Psychology, 94*, 1103–1127. doi:10.1037/a0016172

Clarke, S., & Robertson, I. T. (2005). A meta-analytic review of the Big Five personality factors and accident involvement in occupational and non-occupational settings. *Journal of Occupational and Organizational Psychology, 78*, 355–376. doi:10.1348/096317905X26183

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*, 1241–1255. doi:10.1037/0021-9010.90.6.1241

Dalal, R. S., Lam, H., Weiss, H. M., Welch, E. R., & Hulin, C. L. (2009). A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counterproductivity associations, and dynamic relationships with affect and overall job performance. *Academy of Management Journal, 52*, 1051–1066. doi:10.5465/AMJ.2009.44636148

Dilchert, S., Ones, D. S., Davis, R. D., & Rostow, C. D. (2007). Cognitive ability predicts objectively measured counterproductive work behaviors. *Journal of Applied Psychology, 92*, 616–627. doi:10.1037/0021-9010.92.3.616

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57. doi:10.1037/0021-9010.91.1.40

Duffy, M. K., Ganster, D., & Pagon, M. (2002). Social undermining in the workplace. *Academy of Management Journal, 45*, 331–351. doi:10.2307/3069350

Einarsen, S., Aasland, M. S., & Skogstad, A. (2007). Destructive leadership behaviour: A definition and conceptual model. *Leadership Quarterly, 18*, 207–216. doi:10.1016/j.leaqua.2007.03.002

Fox, S., Spector, P. E., Goh, A., & Bruursema, K. (2007). Does your coworker know what you're doing? Convergence of self- and peer-reports of counterproductive work behavior. *International Journal of Stress Management, 14*, 41–60. doi:10.1037/1072-5245.14.1.41

Giacalone, R. A., & Greenberg, J. (Eds.). (1997). *Antisocial behavior in organizations*. Thousand Oaks, CA: Sage.

Gough, H. G. (1971). The assessment of wayward impulse by means of the Personnel Reaction Blank. *Personnel Psychology, 24*, 669–677. doi:10.1111/j.1744-6570.1971.tb00380.x

Gruys, M. L., & Sackett, P. R. (2003). Investigating the dimensionality of counterproductive work behavior. *International Journal of Selection and Assessment, 11*, 30–42. doi:10.1111/1468-2389.00224

Henig, M. S. (1927). Intelligence and safety. *Journal of Educational Research, 16*, 81–87.

Hershcovis, M. S., Turner, N., Barling, J., Arnold, K. A., Dupré, K. E., Inness, M., . . . Sivanathan, N. (2007). Predicting workplace aggression: A meta-analysis. *Journal of Applied Psychology, 92*, 228–238. doi:10.1037/0021-9010.92.1.228

Hoffman, B. J., & Dilchert, S. (2012). A review of citizenship and counterproductive behaviors in organizational decision-making. In N. Schmitt (Ed.), *Oxford handbook of personnel assessment and selection* (pp. 543–569). New York, NY: Oxford University Press.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis*. Thousand Oaks, CA: Sage.

Ilies, R., Fulmer, I. S., Spitzmuller, M., & Johnson, M. D. (2009). Personality and citizenship behavior: The mediating role of job satisfaction. *Journal of Applied Psychology, 94*, 945–959. doi:10.1037/a0013329

Kaplan, S., Bradley, J. C., Luchman, J. N., & Haynes, D. (2009). On the role of positive and negative affectivity in job performance: A meta-analytic investigation. *Journal of Applied Psychology, 94*, 162–176. doi:10.1037/a0013115

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *Journal of Applied Psychology, 95*, 1–31. doi:10.1037/a0017103

Kornhauser, A. W., & Sharp, A. A. (1932). Employee attitudes: Suggestions from a study in a factory. *Personnel Journal, 10*, 393–404.

Koslowsky, M., Sagie, A., Krausz, M., & Singer, A. D. (1997). Correlates of employee lateness: Some theoretical considerations. *Journal of Applied Psychology, 82*, 79–88. doi:10.1037/0021-9010.82.1.79

Lau, V. C. S., Au, W. T., & Ho, J. M. C. (2003). A qualitative and quantitative review of antecedents of counterproductive behavior in organizations. *Journal of Business and Psychology, 18*, 73–99.

Leymann, H. (1996). The content and development of mobbing at work. *European Journal of Work and Organizational Psychology, 5*, 165–184. doi:10.1080/13594329608414853

Marcus, B., & Schuler, H. (2004). Antecedents of counterproductive behavior at work: A general perspective. *Journal of Applied Psychology, 89*, 647–660. doi:10.1037/0021-9010.89.4.647

Martin, L. E., Brock, M. E., Buckley, M. R., & Ketchen, D. J., Jr. (2010). Time banditry: Examining the purloining of time in organizations. *Human Resource Management Review, 20*, 26–34. doi:10.1016/j.hrmr.2009.03.013

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335–354. doi:10.1111/j.1744-6570.1990.tb01562.x

Moberg, D. J. (1997). On employee vice. *Business Ethics Quarterly, 7*, 41–60. doi:10.2307/3857208

Motley, R. (1926). Lateness of plant employees: A study of causes and cures. *Journal of Personnel Research, 5*, 1–3.

Ng, T. W. H., & Feldman, D. C. (2008). The relationship of age to ten dimensions of job performance. *Journal of Applied Psychology, 93*, 392–423. doi:10.1037/0021-9010.93.2.392

Ng, T. W. H., & Feldman, D. C. (2010). Organizational tenure and job performance. *Journal of Management, 36*, 1220–1250. doi:10.1177/0149206309359809

O'Leary-Kelly, A. M., Griffin, R. W., & Glew, D. J. (1996). Organization-motivated aggression: A research framework. *Academy of Management Review, 21*, 225–253.

Ones, D. S. (1993). *The construct validity of integrity tests*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Ones, D. S., Connelly, B. S., Viswesvaran, C., & Salgado, J. F. (2008). *Counterproductive work behaviors*. Unpublished manuscript.

Ones, D. S., & Viswesvaran, C. (2001a). Integrity tests and other criterion-focused occupational personality scales (COPS) used in personnel selection. *International Journal of Selection and Assessment, 9*, 31–39. doi:10.1111/1468-2389.00161

Ones, D. S., & Viswesvaran, C. (2001b). Personality at work: Criterion-focused occupational personality scales used in personnel selection. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace* (pp. 63–92). Washington, DC: American Psychological Association. doi:10.1037/10434-003

Ones, D. S., & Viswesvaran, C. (2003). Personality and counterproductive work behaviors. In A. Sagie, S. Stashevsky, & M. Koslowsky (Eds.), *Misbehavior and dysfunctional attitudes in organizations* (pp. 211–249). Hampshire, England: Palgrave Macmillan.

Ones, D. S., & Viswesvaran, C. (2008). Costumer service scales: Criterion-related, construct, and incremental validity evidence. In J. Deller (Ed.), *Research contributions to personality at work* (pp. 19–46). Mering, Germany: Hampp.

Ones, D. S., & Viswesvaran, C. (2011). Individual differences at work. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *The Wiley-Blackwell handbook of personality and individual differences* (pp. 379–407). Chichester, England: Blackwell.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Personality at work: Raising awareness and correcting misconceptions. *Human Performance, 18*, 389–404. doi:10.1207/s15327043hup1804_5

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679–703. doi:10.1037/0021-9010.78.4.679

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2003). Personality and absenteeism: A meta-analysis of integrity tests. *European Journal of Personality, 17*(Suppl. 1), S19–S38. doi:10.1002/per.487

Oppler, S. H., McCloy, R. A., & Campbell, J. P. (2001). The prediction of supervisory and leadership performance. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 389–409). Mahwah, NJ: Erlbaum.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Heath.

Puffer, S. M. (1987). Prosocial behavior, noncompliant behavior, and work performance among commission

salespeople. *Journal of Applied Psychology, 72*, 615–621. doi:10.1037/0021-9010.72.4.615

Roberts, B. W., Harms, P. D., Caspi, A., & Moffitt, T. E. (2007). Predicting the counterproductive employee in a child-to-adult prospective study. *Journal of Applied Psychology, 92*, 1427–1436. doi:10.1037/0021-9010.92.5.1427

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38*, 555–572. doi:10.2307/256693

Rotundo, M., & Spector, P. E. (2010). Counterproductive work behavior and withdrawal. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 489–511). New York, NY: Routledge/Taylor & Francis.

Sackett, P. R. (2002). The structure of counterproductive work behaviors: Dimensionality and relationships with facets of job performance. *International Journal of Selection and Assessment, 10*, 5–11. doi:10.1111/1468-2389.00189

Sackett, P. R., & DeVore, C. J. (2001). Counterproductive behaviors at work. In N. Anderson, D. S. Ones, H. Sinangil Kepir, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology: Vol. 1. Personnel psychology* (pp. 145–164). London, England: Sage.

Salgado, J. F. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment, 10*, 117–125. doi:10.1111/1468-2389.00198

Skarlicki, D. P., & Folger, R. (1997). Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. *Journal of Applied Psychology, 82*, 434–443. doi:10.1037/0021-9010.82.3.434

Spector, P. E. (1975). Relationships of organizational frustration with reported behavioral reactions of employees. *Journal of Applied Psychology, 60*, 635–637. doi:10.1037/h0077157

Spector, P. E., & Fox, S. (2005). The stressor-emotion model of counterproductive work behavior. In S. Fox & P. E. Spector (Eds.), *Counterproductive work behavior: Investigations of actors and targets* (pp. 151–174). Washington, DC: American Psychological Association. doi:10.1037/10893-007

Stewart, S. M., Bing, M. N., Davison, H. K., Woehr, D. J., & McIntyre, M. D. (2009). In the eyes of the beholder: A non-self-report measure of workplace deviance. *Journal of Applied Psychology, 94*, 207–215. doi:10.1037/a0012605

Tepper, B. J. (2000). Consequences of abusive supervision. *Academy of Management Journal, 43*, 178–190. doi:10.2307/1556375

U.S. Bureau of Labor Statistics. (2006). *Survey of workplace violence prevention, 2005*. Washington, DC: U.S. Department of Labor.

Vardi, Y., & Wiener, Y. (1996). Misbehavior in organizations: A motivational framework. *Organization Science, 7*, 151–165. doi:10.1287/orsc.7.2.151

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment, 8*, 216–226. doi:10.1111/1468-2389.00151

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574. doi:10.1037/0021-9010.81.5.557

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131. doi:10.1037/0021-9010.90.1.108

Wheeler, H. N. (1976). Punishment theory and industrial discipline. *Industrial Relations, 15*, 235–243. doi:10.1111/j.1468-232X.1976.tb01120.x

Zapf, D., Einarsen, S., Hoel, H., & Vartia, M. (2003). Empirical findings on bullying in the workplace. In S. Einarsen, H. Hoel, D. Zapf, & C. D. Cooper (Eds.), *Bullying and emotional abuse in the workplace: International perspectives in research and practice* (pp. 103–126). London, England: Taylor & Francis.

CHAPTER 36

STEREOTYPE THREAT IN WORKPLACE ASSESSMENTS

Ann Marie Ryan and Paul R. Sackett

Given the widespread use of assessment tools for workplace decision making as well as findings regarding group differences in performance on those assessments (see Sackett, Schmitt, Ellingson, & Kabin, 2001, for a review), understanding environmental factors that might differentially affect performance is an important concern for those who develop and administer assessments in the workplace. One such potential factor is *stereotype threat,* the inhibition of performance on a task because of concern about confirming a stereotype. The aim of this chapter is to discuss the applicability of existing research on stereotype threat to workplace assessments and the implications of that research for the design and use of assessment tools in hiring, training, performance evaluation, and other work-related contexts.

Sackett and Ryan (2011) differentiated between two contexts in which stereotype threat is invoked. The first context, and the focus of this chapter, is settings in which the proposition is that threat has an artifactual effect on test scores (i.e., test takers do not demonstrate their true standing on the construct of interest because of threat). The second context is settings involving intervention aimed at true change in the construct of interest. For example, Good, Aronson, and Inzlicht (2003) described an intervention over the course of a school year that imparted a position that intelligence was malleable to assess whether it led to higher achievement among seventh graders, as indexed by performance on statewide tests given at the end of the school year. Although the study involved a high-stakes test (e.g., used to influence promotion decisions), the argument was not that threat prevented students from demonstrating their ability when tested but rather that threat interfered with learning throughout the year, and the intervention helped remove those barriers. Again, our focus is on stereotype threat as a potential impediment to assessing current standing on a construct of interest, because it is in that context that stereotype threat is a potential impediment to unbiased assessment.

This chapter starts with a brief discussion of the basic tenets of stereotype threat theory. Whether the particular contextual features of workplace assessments meet the theoretical requirements for the phenomenon to occur is noted. A more detailed summary of the very few studies conducted with actual workplace assessments is presented, and the need for and challenges in conducting research on this topic in workplace contexts is noted. Common misinterpretations of research findings on stereotype threat are discussed. The chapter concludes with thoughts on the applicability of threat reduction strategies for workplace assessment contexts.

THEORETICAL BOUNDARIES OF STEREOTYPE THREAT AND APPLICABILITY TO WORKPLACE ASSESSMENTS

Stereotype threat affects performance when an individual underperforms on a task because of concern about confirming a negative stereotype about a group with which he or she identifies (Steele &

DOI: 10.1037/14047-036
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

Aronson, 1995). Stereotype threat is typically induced in research by manipulating either stereotype salience or identity salience (Campbell & Collaer, 2009). For example, noting typical differences in math scores of men and women immediately before administering a math test would make that stereotype salient; making individuals self-identify their gender on a demographics form makes that identity salient. For stereotype threat effects to occur, a number of conditions need to be present (Steele, 1997, 2010); each of these conditions and their applicability to the typical workplace assessment context is described in this section.

1. *A consistent stereotype exists, and group members are aware of its existence.* One can think of many types of workplace assessments for which a consistent stereotype exists: beliefs regarding ethnic minorities and cognitive ability, women and math, age and technology, women and physical strength, and women and leadership. Note that the presence of a stereotype may reflect a true state of affairs. For example, reports that on average men have greater physical strength than women is not controversial. The question of interest is whether the observed difference provides a distorted estimate of the true difference because of the presence of stereotype threat.

One should not automatically assume that test takers are aware of the stereotype's existence. For example, younger students in France believe girls are better in math than boys (Martinot & Desert, 2007); much research on stereotype threat awareness focuses on changes in awareness in young children. However, some evidence of individual differences in stereotype awareness has been found in adults as well. Hall and Carter (1999) illustrated that the ability to accurately note gender differences is an individual difference, although across five samples accuracy was quite high, suggesting that most people are aware of gender stereotypes. Others have also documented the pervasiveness of knowledge of many stereotypes (Augoustinos, Innes, & Ahrens, 1994; Devine, 1989; Lepore & Brown, 1997). Thus, this condition would generally be met for a number of different types of workplace assessments and for individuals of a number of different social categories. Further research on stereotype awareness among applicants for common workplace assessments would shed light on how often this condition is met.

2. *The task must be seen as diagnostic.* Assessments in the workplace are used for evaluative purposes because they serve as the means to evaluate individual suitability for hire or promotion or advancement from training. Sometimes assessments are used in a developmental context (e.g., giving feedback on leadership skills), in which they are seen as diagnostic of some underlying skill, ability, knowledge, or personal characteristic. Once again, evidence has been found of variability across individuals in perceptions of the diagnosticity of assessments. For example, in a meta-analysis of applicant reactions to testing procedures across 17 countries, Anderson, Salgado, and Hulsheger (2010) found variability across individuals in perceptions of the validity of commonly used selection tools. We wonder too whether the boundary condition here was beliefs about the appropriateness of using something for diagnosis or the fact that something is used to evaluate or judge, regardless of whether one thinks that it is effective for that aim. Further research on what the precise nature of the boundary condition is would be useful (e.g., if individuals do not see a test as high in predictive validity, are they less susceptible to stereotype threat even if the test is used to evaluate them?).

3. *The stereotype must be relevant to the individual during a situation in which the individual is at risk of confirming the stereotype.* Relevance of a stereotype will be affected by what the individual believes is being assessed. For example, a stereotype regarding gender and leadership will not induce stereotype threat if a female test taker does not recognize the test as a measure of leadership ability. For many assessment tools, it may be obvious to the test taker what the aim is, and most organizations do try to be explicit about what constructs are assessed. However, there are cases in which test takers may not understand the real purpose of an assessment (e.g., an interview question that is very general, such as "Tell me about yourself") or in which a measure is not particularly face valid. In those cases, even if a stereotype exists regarding the performance of the individual's social group, threat will not be experienced because the stereotype is not seen as relevant to the assessment.

For the many workplace assessments in which individuals know the construct being assessed, are stereotypes seen as relevant? Typical stereotype threat research studies involve some manipulation to make the stereotype salient, but the actual saliency of stereotypes in real-world assessment contexts likely varies. Many stereotype threat studies have relied on a blatant presentation of a stereotype before testing to induce the effect, such as stating that one group is expected to perform more poorly, something that would not occur in any workplace assessment settings of which we are aware (Sackett, Hardison, & Cullen, 2004). Therefore, it is important to consider the research studies involving more subtle cues—such as manipulations of the testing environment to increase identity salience—as a more realistic simulation of what might occur in workplace testing. Nguyen and Ryan (2008) noted that effect sizes of stereotype threat on cognitive tests do appear to differ depending on cue type as well as group stereotyped, but subtle cues do produce moderate effects in lab settings. Are subtle cues present in workplace assessment contexts? Being asked to fill out a demographic questionnaire before taking a test or taking the assessment in a room in which the majority are of one ethnic, racial, or gender group are common subtle cue manipulations that may be present in many workplace assessment contexts.

Steele (1997) has argued that stereotype threat is "in the air" in that one need not manipulate anything to make stereotypes relevant in high-stakes assessment contexts because their relevance is pervasive for members of the social groups to which they apply. However, given the wide variability in what is assessed and how it is assessed in today's workplace tools, it is important to measure whether a stereotype is seen as linked to a particular tool rather than assuming this occurs; organizational psychologists could contribute much to the understanding of boundaries to effects by addressing this issue for common assessment methods. That is, individuals must be aware of the stereotype's existence, must have a sense of what construct the assessment is meant to measure, and see those as linked in that context. In sum, although this condition may often be met for several popular types of assessments, it is not something that should be universally assumed in the typical uses of workplace assessments.

4. *The task must be "at the frontier of a person's skills" (Steele, 2010, p. 109); in other words, the assessment needs to be a highly difficult one.* In his initial stereotype threat experiments, Steele (2010) noted that on average test takers got only 30% of the items correct—the tests were very difficult even for a highly select group, Stanford students. Nguyen and Ryan (2008) also noted that test difficulty moderates stereotype threat effects on cognitive ability tests, with stronger effects for highly difficult rather than moderately difficult tests.

Are workplace assessments of high difficulty? One challenge in answering this question is that the theoretical proposition has to do with difficulty in terms of the individual's skill level, not as a property of the test for a group of test takers. For example, some training assessments and some selection screening tools have a low cut score in that their aim is to screen out the truly unqualified. However, the fact that the cut score is low is likely not known to the applicant, whose judgment of difficulty is likely based on the perceived proportion of items answered correctly. Some measures are designed for broad use (e.g., a commercially published cognitive ability test), and the difficulty experienced is likely to vary widely across jobs that attract applicant pools differing in ability. Other job assessments are tailored to a specific job, and care is taken to include no items more difficult than required for the job. Indeed, court cases have revolved around employees' testing for skills not required or at a level of difficulty beyond the job requirements. For example, the *Lanning v. Southeastern Pennsylvania Transportation Authority* (1999, 2002) case required setting cutoff scores at the minimum qualification level rather than as high as an organization might desire for business purposes. Although this case was not about the difficulty of the test and other guidelines and cases have implied that employers simply need a reasonable business rationale for setting cut scores, it highlights that many employers do not give very difficult assessments but design difficulty levels around minimal rather than maximum job requirements. In sum, most assessments will not be

perceived as highly difficult for those with the skills needed for the job but may be difficult for applicants lacking those skills. Thus, it is possible that some applicants may experience threat, but those who do are those lacking the needed skills—which suggests that a question more nuanced than "Do applicants experience threat?" is needed. The question of interest is "Do applicants with the skill level that would result in their selection absent threat experience threat?"

Steele (2010) noted that if a task is within an individual's skill level, he or she will not experience frustration but will perform well in the presence of a stereotype. Employers may encourage individuals to apply only for jobs for which they are clearly qualified (i.e., within their skill level) by, for example, having them do self-assessments of job fit and practice assessment items. Moreover, there is the interesting possibility that test-taker perceptions of difficulty may be affected by threat mechanisms. That is, if stereotype threat does draw cognitive resources away from answering test questions (Schmader, Johns, & Forbes, 2008), then it is likely that those test takers will feel the test is more difficult. However, for many workplace assessment contexts and many test takers, this condition of highly difficult may not be met.

5. *The individual must identify with the domain with which the stereotype is linked; that is, the individual must care about being seen as being skilled and capable in that domain.* Steele (2010) referred to those who are strongly identified as the vanguard in an area. Whether this vanguard condition is fully met in workplace assessments is interesting to consider. More often than not, assessments in the workplace are high stakes. That is, assessments typically serve as input into decisions that have important life consequences for the individuals taking them, such as obtaining a job, obtaining training certification, or being marked as having high potential for advancement. Workplace assessments are given in contexts in which individuals are seeking highly desired outcomes and have a very high motivation to perform well; thus, individuals in these contexts do care about being viewed as competent.

However, typical test takers in workplace settings will vary substantially in domain identification. Workplace assessments cover a wide variety of constructs (general cognitive ability, specific job knowledge, personality characteristics, work styles). Some of these constructs may be domains with which individuals strongly identify (e.g., knowledge of psychology for individuals seeking a psychology license), but many others will be ones in which much greater variability in identification for a given pool of test takers will occur (e.g., conscientiousness for a plant technician position). To illustrate, individuals might need to take a basic math skills test for an entry-level customer service representative position in which they will be handling cash and counting inventory. If asked, a sizable portion of the applicant pool will likely report being not highly math identified (i.e., they are not math majors or individuals pursuing careers requiring high math skills). Lab studies in which stereotype threat effects are produced involve preselecting individuals who care about the domain; for many job skills that are the basis for screening instruments, most applicants will not feel the level of identification present in these lab samples.

Logel, Iserman, Davies, Quinn, and Spencer (2009) noted that in real-world settings, even those not strongly identified with the domain will be motivated to do well because of the consequences associated with performance. That is, individuals' desire to perform well on workplace assessments may be both intrinsically and extrinsically motivated in some cases, but only extrinsically motivated (i.e., I must do well to get this job) in others. However, motivation and domain identification are hardly synonymous, raising questions about whether the domain identification condition is met for many assessees in typical workplace assessment contexts. Further research that separately examines the independent effects of domain identification and test-taker motivation would be useful.

6. *Some level of identification with the stereotyped group must exist.* Steele (2010) described stereotype threat as a stigma-related contingency of a specific identity; therefore, individuals must see the identity as applicable to themselves. For example, if a 55-year-old man is well aware of stereotypes regarding age and technology and is taking an assessment evaluating his technological skills but does not view himself as old, threat is less likely to be induced,

because individuals are not threatened by stereotypes that they do not see as applying to themselves. If the stereotype is not seen as applicable (e.g., Whites and cognitive ability tests), performance will be either unaffected or, in some cases, an increase can occur because of stereotype lift (Walton & Cohen, 2003). Presumably, many individuals will have some level of identification with their gender, ethnic, and other social categories; what level of identification is required for stereotype effects to occur remains an open question. Schmader (2002) demonstrated that women lower in gender identification were less vulnerable to stereotype effects on math tests.

In summary, although many workplace assessment contexts will meet some of the boundary conditions of the theory for many test takers (e.g., awareness of a common stereotype), considerable variability is likely in whether other conditions are met in a particular assessment context for various test takers (e.g., highly difficult, high domain identification). Lab studies are purposely designed to ensure the conditions to produce an effect are present—individuals can be made aware of a stereotype, and its relevance to the task can be pointed out to them directly. Individuals can be preselected for study participation on the basis of their levels of domain identification, and tests can be selected that are highly difficult for that group. In workplace contexts, greater variability across contexts and people in these theoretically required conditions is highly likely. Assuming that stereotype threat automatically occurs when tests are administered for workplace decision making is not in line with the basic tenets of Steele's (2010) theory of stereotype threat.

STEREOTYPE THREAT RESEARCH CONDUCTED IN WORKPLACE SETTINGS

Having established that the conditions for stereotype threat to occur are not always present in workplace contexts, let us consider the question of whether there is any evidence that stereotype threat effects do occur in the workplace. Hundreds of studies on stereotype threat effects have been conducted in lab settings. However, only a handful of studies have specifically sought to examine stereotype threat with regard to workplace assessments. This small body of research has some clear limitations.

First, several studies have looked at simulated applicant settings (Mayer & Hanges, 2003; McFarland, Lev-Arey, & Ziegert, 2003; Nguyen, O'Neal, & Ryan 2003; Ployhart, Ziegert, & McFarland, 2003), in which a lab study participant is asked to role-play an applicant taking an assessment for employment purposes. These studies have generally found little evidence of stereotype threat effects, but there are limitations to interpreting the lack of observed effects because of the absence of a true control (i.e., non diagnostic testing condition; Steele & Davis, 2003). These studies also suffer from concerns that plague many laboratory studies on selection contexts in that research participants are simply not as motivated to succeed as are real-world applicants.

Second, several studies have been focused on educational admission and placement contexts. These studies have had contextual features similar to workplace assessment; these assessments are high-stakes ones given to examinees with a strong interest in obtaining admission to the institution of interest or placement into an advanced course on the basis of evidence of prior achievement. Also, educational tests in the cognitive ability domain are very similar to cognitive ability tests used in the employment setting (Frey & Detterman, 2004). As in the workplace setting, replicating the common laboratory stereotype paradigm in an operational educational admissions setting is difficult, and so two strategies are used. One is to rely on the differential prediction paradigm commonly used to assess predictive bias in the relationship between test scores and criteria. This approach focuses on regression lines relating test scores and criteria for threatened and nonthreatened groups and develops predictions as to the effects that stereotype threat would have on these regression lines if threat was affecting scores. Two studies have used this approach with operational admissions data; neither has found evidence of the pattern of relationships that would be expected if threat were operating (Cullen, Hardison, & Sackett, 2004; Cullen, Waters, & Sackett, 2006).

The other strategy is to rely on the limited opportunities for experimentation or quasi-experimentation.

One study receiving considerable attention (Stricker & Ward, 2004) randomly assigned test takers to report demographic information (i.e., race, gender) either before or after completing a test, a subtle manipulation that had been found to produce threat effects in laboratory settings. Stricker and Ward (2004) looked at several tests and concluded that the pattern of results did not produce evidence of threat effects. Danaher and Crandall (2008) reanalyzed the Stricker and Ward data and argued that there was indeed support for threat effects. Looking at one test of advanced placement calculus, they found that women performed more poorly when asked to report gender before the test, and they projected the magnitude of the effect to the full test-taking population and estimated that threat resulted in the wrongful denial of advanced placement credit to 4,731 women. Sackett and Ryan (2011) took issue with this conclusion, noting that for a second test an effect of roughly comparable magnitude was found, but in the opposite direction (women performed more poorly if asked to report gender after the test) and that for a third test there was no effect in either direction. Thus, only a selective review of the data from Stricker and Ward supports an interpretation of threat. Considering all of the evidence, one would conclude no net effect exists for the subtle intervention in a real test-taking situation. Other studies in operational testing settings include Walker and Bridgeman's (2008) examination of a potential spillover effect (i.e., whether scores on a critical reading subtest differed as a result of the test's following a math, reading, or writing subtest). Walker and Bridgeman concluded that there was no evidence of stereotype threat resulting from the type of test to which examinees were first exposed. Also, Walters, Lee, and Trapani (2004) examined whether having a proctor of the same or a different race or ethnicity or of a different gender affected operational GRE performance. Findings were not consistent with what would be expected if a different-race or different-gender proctor induced stereotype threat. Only a few significant differences were found, and those differences were small effects in the opposite direction than expected.

Moving to studies in the employment domain, one study examined archival applicant data. Kirnan, Alfieri, Bragger, and Harris (2009) examined whether demographic questions asked before or after a cognitive ability test served as a stereotype threat cue. They did not find strong evidence of stereotype threat effects; however, they noted a number of weaknesses in their design (i.e., differences in test difficulty and potentially in applicant pool quality across conditions) that make interpretations difficult.

Also, one field study examined self-reported stereotype threat and found little evidence of an effect on performance in a promotion context (Chung, Ehrhart, Ehrhart, Hattrup, & Solamon, 2010). However, in his writing, Steele (1977) has made it clear that one need not be consciously aware of threat to experience its effects (Steele, 1997), so the conclusions one can reach from this study are also limited.

Recently, Meyer and Melchers (2010) examined stereotype threat in performance in computer-simulated problem-solving tasks called *microworlds*. The microworld task was seen as stereotypically male linked in that it involved computer-based administration and was presented as being diagnostic of managerial performance, and ample research has indicated negative stereotypes of women's managerial competence (Duehr & Bono, 2006). In Study 3 of their article, they examined applicants for entry-level management positions at a bank who participated in the microworld at an assessment center. They considered gender composition of the group being assessed as an index of experienced threat and found that women who were solo in the group performed more poorly in the microworld than those in groups with more women (accounting for an estimated 5.8% of variance in performance).

However, Meyer and Melchers (2010) also considered a group discussion exercise as a task that was not stereotyped, and it showed no effects of group composition on performance in the discussion. They argued that the discussion was not a stereotyped task because it was a verbal discussion, and there are no negative stereotypes about women's verbal abilities. However, all of the exercises were in the context of a managerial assessment center, and Meyer and Melchers (2010) offered general stereotypes about managerial jobs ("think manager, think male") as the basis for their expectations for gender differences in microworld performance. So why

threat effects would be observed for one exercise but not for others is not clear.

In sum, really very little research exists specifically on the topic of stereotype threat effects in workplace assessment contexts, and the only study to demonstrate effects is the unpublished Meyer and Melchers (2010) piece. Two questions seem important to address: (a) Why can the large body of research on stereotype threat produced in lab settings not be viewed as generalizing to the workplace assessment context, and (b) why are there not more studies conducted in workplace assessment contexts?

The answer to the first question remains a source of debate. In this chapter, we have outlined a number of key conditions noted by leading stereotype threat theorists as being necessary for threat effects to occur, and we have illustrated that although some are most likely present in most workplace assessment contexts, others are likely to be absent from many applied settings. This does not mean that stereotype threat effects do not occur with workplace assessments, but rather that assuming they are ubiquitous would not be in keeping with the tenets of the theory. Generalizing from lab to field settings involves moving from the question of "Can this happen?" to "Does this happen?" Although the hundreds of lab studies with student participants show that one can indeed produce stereotype threat effects, the differences between these studies and prototypical workplace assessments in terms of stereotype salience, task difficulty, test-taker domain identification, and other aspects still leave the question of "Does this happen?" unanswered.

The answer to the second question of why there is not more research on workplace assessments resides in the methodologies used in the laboratory to produce stereotype threat effects. As noted earlier, studies manipulate either stereotype salience or identity salience, by either increasing them or reducing them. Typical manipulations of stereotype salience (telling test takers that those of one group perform more poorly on a test right before administering it or telling test takers there are no differences between groups) are not going to occur in workplace settings. Indeed, the latter would be highly problematic in a legal and ethical sense when it is in fact false (i.e., there are group differences in cognitive ability test scores). As discussed in the next section, manipulating identity salience is also challenging in workplace contexts. Indeed, any type of experimental manipulation that is thought to affect test scores in a high-stakes context would be ethically (and likely legally) problematic (Nguyen & Ryan, 2008), which leaves researchers with examining naturally occurring quasi-experiments (e.g., a change in testing procedure; Kirnan et al., 2009) or using observational studies as the only feasible research designs, which certainly pose more challenges to making causal inferences regarding stereotype threat effects.

APPLICABILITY OF STEREOTYPE THREAT REDUCTION STRATEGIES TO WORKPLACE ASSESSMENTS

One question that arises from workplace assessment users is, "What can one do to reduce the likelihood of stereotype threat effects?" That is, from a practical standpoint, whether one knows for sure an effect is occurring, what types of strategies might be used to lessen its likelihood?

A strong desire may exist to see reducing stereotype threat as a solution to the problem of adverse impact (i.e., disproportionate hiring rates) associated with many common selection assessments tools. As individuals attempt to draw inferences regarding the meaning of stereotype threat research for use of assessments in workplace contexts, care must be taken not to make some common misinterpretations of this research. That is, stereotype threat is a within-group effect and as such should not be presented automatically as an explanation for between-groups differences (see Sackett, 2003, and Sackett et al., 2004, for examples of this misinterpretation). Research on eliminating stereotype threat has been interpreted as demonstrating an elimination of group differences, but it is not the case because individuals in different groups are statistically equated on ability measures in these studies. That is, removing stereotype threat still leaves sizable group differences in performance. For example, even if a particular use of a cognitive ability test for selection was a context that would meet all the conditions noted earlier for stereotype threat to occur

for minority test takers, removing or reducing threat in that context would not lead to a total elimination of group differences in performance on the cognitive ability measure.

With that very important caveat in mind, one can look to the laboratory research on stereotype threat removal as to what strategies have been promoted as effective and examine whether they can be used in the typical workplace assessment context.

1. *Nullify the stereotype (Spencer, Steele, & Quinn, 1999).* One of the most widely studied methods for reducing stereotype threat is the nullification of the stereotype—telling individuals that it is not true. As noted earlier, nullification is not possible when it is contrary to evidence. Many (but not all) tests used in employment settings show sizable group differences, and to state that differences do not exist does not refute a stereotype but rather provides false information to test takers. Although this strategy is touted as a simple one for stereotype threat researchers to use, it is not an ethical one to use outside of an artificial lab setting (Campbell & Collaer, 2009).

A less potent manipulation than nullifying a stereotype is finding ways to "decouple" it from the assessment or make it less salient in that context. For example, Steele (2010) reported telling test takers that although the stereotype does exist, it is not true of that particular test. Once again, the challenge here would be whether an employer could indeed make that statement about his or her particular assessment tool. Given the ubiquity of group differences on cognitive ability tests (Sackett et al., 2001), the ability to use decoupling strategies for many workplace assessments is unlikely.

2. *Describe the assessment as nondiagnostic (Alter, Aronson, Darley, Rodriguez, & Ruble, 2010; Steele & Aronson, 1995).* Another commonly used removal strategy in stereotype threat research is to tell individuals that the task in which they are engaging is not diagnostic of ability in an area, often describing it simply as a measure of problem solving. This strategy is also entirely infeasible in workplace assessment contexts, where stating that something will not be used to diagnose skills and abilities when that is the exact purpose of the assessment is ethically inappropriate.

3. *Test in a same-group environment (Inzlicht & Ben-Zeev, 2000).* We see this strategy as also falling into the infeasible category. Organizations cannot always control the composition of a group that shows up for testing, and test takers might question being separated by group. Indeed, because of legal prohibitions against differential treatment of applicants, an organization might incur a lawsuit if it assigned applicants to testing settings on the basis of ethnicity, race, or gender.

The strong movement toward computerized testing and in particular unproctored Internet testing suggests that test takers will have more control over their testing environment and may be much more likely to be testing alone, avoiding this as a potential contributor to effects.

4. *Affirm the self (e.g., write several paragraphs about one's most important values; Cohen, Garcia, Apfel, & Master, 2006).* Using self-affirmation as a strategy in a workplace assessment context has several real limitations. Employers are generally reluctant to ask individuals for any type of information that they are not directly going to use in decision making, for reasons of making the process as efficient as possible in terms of time costs on the part of administrators and especially test takers, limiting legal exposure, and also not misleading individuals as to the basis for decision making. Ample research has suggested that individuals who are being assessed for decision-making purposes will seek to create a positive impression (Bolino, Kacmar, Turnley, & Gilstrap, 2008; McFarland, Ryan, & Kriska, 2003); self-affirmation in an employment context will likely be viewed through a desire to impression manage. Using a self-affirmation exercise as a formal part of an assessment process would be problematic for all of these reasons.

5. *Frame the characteristic being assessed as malleable (Aronson, Fried, & Good, 2002).* Researchers have suggested that intervening to get test takers to adopt a more incremental than entity perspective can eliminate stereotype threat effects. At first blush, this intervention may seem innocuous for workplace assessments. However, some characteristics are not easily trained or changed and to suggest they are would be unethical. Most of the characteristics on which employers screen job applicants are assumed

to be fairly stable individual differences (e.g., cognitive ability, conscientiousness); however, some change in fluid and crystallized intelligence is known to occur in one's early working years (McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002), and domain-specific knowledge can increase into late adulthood (see Reeve & Hakel, 2000, for a review). With regard to personality, Roberts and Mroczek (2008) recently noted that individuals tend to increase in agreeableness, emotional stability, and conscientiousness from young adulthood to middle age. Thus, although these characteristics do change some during one's working years, their relative stability has been well documented (Ackerman & Humphreys, 1990; Blonigen, Carlson, Hicks, Krueger, & Iacono, 2008). Moreover, suggesting that something is highly malleable seems to imply that one need not screen on it, raising concerns about why an employer would be evaluating individuals on something that the employer could easily train. The legal precedent is considerable regarding the need for employers to avoid screening on skills that can easily be trained if adverse impact results from their assessment. Professional guidelines have noted that one should select on knowledge, skill, ability, or other characteristics that one needs to know without training (Society for Industrial and Organizational Psychology, 2003).

The use of this strategy might be much more appropriate in settings in which the assessment is solely for developmental purposes (e.g., assessing leadership skills as part of a leader development program); indeed, the presumption in those cases is that the assessments are being made on characteristics that are malleable.

6. *Increase accessibility of social identities associated with positive stereotypes in the domain (Rydell, McConnell, & Beilock, 2009).* Some of the strategies suggested as ways to increase the accessibility of positive stereotypes (e.g., describe overlapping traits of men and women) are unlikely to occur in workplace assessment contexts as a result of both time constraints and concern over applicant perceptions of relevance to the assessment purpose. However, subtle statements can be made in the testing materials that draw attention to other identities (e.g., noting the experience level of job seekers). One challenge that Rydell et al. (2009) noted is that positive stereotype manipulations that work for one group may not work for others, or may even increase threat for others (e.g., making salient education levels of college-educated minorities as associated with success may create threat for those lacking educational credentials if variability in education levels occurs in the applicant pool). Thus, use of this strategy should be approached with caution.

7. *Teach individuals about stereotype threat (Johns, Schmader, & Martens, 2005).* In this form of intervention, the stereotype is described and individuals are told that

> it's important to keep in mind that if you are feeling anxious while taking this test, this anxiety could be the result of these negative stereotypes that are widely known in society and have nothing to do with your actual ability to do well on the test. (Johns et al., 2005, p. 176)

One might be hard pressed to imagine organizational legal teams allowing even a mention of a group difference as part of a testing orientation, let alone a tutorial on stereotype threat effects. However, at a broader societal level rather than within a specific testing context, such teaching could occur (e.g., workshops for job seekers).

8. *Expose individuals to positive role models (Marx & Roman, 2002; McIntyre, Paulson, & Lord, 2003) and positive social comparison information (Marx, Stapel, & Muller, 2005).* Typical interventions along these lines include sharing success stories of one's identity group in the stereotyped domain just before the assessment. Many organizations might provide such exposure by including successful minority and female employees as examples in their recruitment literature, as recruiters, and in interactions on site visits (Avery, 2003; Avery, Hernandez, & Hebl, 2004). However, this exposure is not typically directly connected to the assessment context. Although direct positive social comparison information would not be provided for only some groups in an assessment context (i.e., one would want consistency rather than differential treatment), there may be opportunities elsewhere in the recruiting process (e.g., in Web profiles of employees) to highlight positive role models.

9. *Reframe the assessment (e.g., Stone, Lynch, Sjomeling, & Darley, 1999).* Some form of the strategy of reframing the assessment may be possible in many workplace assessment contexts. For example, organizations can be mindful of how the assessment is described. In our experience, few employers state they are examining intelligence, but they often talk about capacity to learn or problem-solving ability when administering cognitive ability tests, and thus some of this framing already occurs. Others (Kirnan et al., 2009) have noted that explaining the test, its uses, and general fairness may increase trust in the testing situation. These all are standard good practices that should be followed in all assessment contexts.

In sum, many of the means of stereotype threat removal or reduction studied in the lab are not feasible or appropriate for workplace assessment contexts. Key stereotype threat researchers have also noted that suggested methods of reducing threat are difficult to translate into practical interventions (Johns et al., 2005). Those methods that appear most feasible (exposure to positive role models in recruitment and reframing assessments) are already carried out by many larger organizations and high-quality test programs, not because of a known connection to stereotype threat reduction but as a generally good recruitment and selection practice.

Despite a recent strong focus on delineating the process mechanisms of stereotype threat (e.g., off-task thinking, anxiety; Schmader et al., 2008), these mechanisms have, interestingly, not been the focus of discussions on reduction of stereotype threat. Presumably, researchers have aimed to prevent stereotype threat from occurring at all, rather than working with individuals to reduce the mechanisms by which lowered performance occurs. In recent years, the employment testing community has shown a keen interest in making the applicant testing experience a positive one (Ryan & Ployhart, 2000), so there is likely great applied interest and willingness to consider assessment framing, instructions, orientation materials, and even candidate preparation programs to reduce anxiety and other sources of construct irrelevant variance in test scores.

Steele (2010) recently suggested that one does not have to change all setting cues to remove stereotype threat, just enough for a critical degree of "identity safety." Perhaps employers who provide an assessment context in which they use valid tools, describe the purpose of the assessment and the use of the information, and ensure that test takers have sufficient familiarity with the assessment and a comfortable assessment environment are ones who are already increasing the safety of the assessment process.

CONCLUSION

This chapter has highlighted that stereotype threat is a well-established phenomenon in laboratory contexts but its assumed pervasiveness in workplace assessment contexts deserves strong scrutiny on both conceptual and empirical grounds. Moreover, methods for removal of stereotype threat effects are touted without consideration of the ethical, legal, and practical constraints of workplace assessment that render the use of many of these methods impossible. What has come to be viewed as a pervasive cause (stereotype threat) of a widespread societal problem (large group differences in test scores) that can be easily remedied (through threat removal strategies) has in reality not yet been established as pervasive, as a cause, or as easily remedied in workplace assessment contexts.

References

Ackerman, P. L., & Humphreys, L. G. (1990). Individual differences theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 223–282). Palo Alto, CA: Consulting Psychologists Press.

Alter, A. L., Aronson, J., Darley, J. M., Rodriguez, C., & Ruble, D. N. (2010). Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. *Journal of Experimental Social Psychology, 46*, 166–171. doi:10.1016/j.jesp.2009.09.014

Anderson, N., Salgado, J. F., & Hulsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment, 18*, 291–304. doi:10.1111/j.1468-2389.2010.00512.x

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence.

Journal of Experimental Social Psychology, 38, 113–125. doi:10.1006/jesp.2001.1491

Augoustinos, M., Innes, J. M., & Ahrens, C. (1994). Stereotypes and prejudice: The Australian experience. *British Journal of Social Psychology, 33*, 125–141. doi:10.1111/j.2044-8309.1994.tb01014.x

Avery, D. R. (2003). Reactions to diversity in recruitment advertising—Are differences black and white? *Journal of Applied Psychology, 88*, 672–679. doi:10.1037/0021-9010.88.4.672

Avery, D. R., Hernandez, M., & Hebl, M. R. (2004). Who's watching the race? Racial salience in recruitment advertising. *Journal of Applied Social Psychology, 34*, 146–161. doi:10.1111/j.1559-1816.2004.tb02541.x

Blonigen, D. M., Carlson, M. D., Hicks, B. M., Krueger, R. F., & Iacono, W. G. (2008). Stability and change in personality traits from late adolescence to early adulthood: A longitudinal twin study. *Journal of Personality, 76*, 229–266. doi:10.1111/j.1467-6494.2007.00485.x

Bolino, M. C., Kacmar, K. M., Turnley, W. H., & Gilstrap, J. B. (2008). A multi-level review of impression management motives and behaviors. *Journal of Management, 34*, 1080–1109. doi:10.1177/0149206308324325

Campbell, S. M., & Collaer, M. L. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychology of Women Quarterly, 33*, 437–444. doi:10.1111/j.1471-6402.2009.01521.x

Chung, B. G., Ehrhart, M. G., Ehrhart, K. H., Hattrup, K., & Solamon, J. (2010). Stereotype threat, state anxiety, and specific self-efficacy as predictors of promotion exam performance. *Group and Organization Management, 35*, 77–107. doi:10.1177/1059601109354839

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313*, 1307–1310. doi:10.1126/science.1128317

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT–grade and ability–job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89*, 220–230. doi:10.1037/0021-9010.89.2.220

Cullen, M. J., Waters, S. D., & Sackett, P. R. (2006). Testing stereotype threat theory predictions for math majors and non-majors by gender. *Human Performance, 19*, 421–440. doi:10.1207/s15327043hup1904_6

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings reexamined: A reply. *Journal of Applied Social Psychology, 34*, 1656–1663.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18. doi:10.1037/0022-3514.56.1.5

Duehr, E. E., & Bono, J. E. (2006). Men, women, and managers: Are stereotypes finally changing? *Personnel Psychology, 59*, 815–846. doi:10.1111/j.1744-6570.2006.00055.x

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15*, 373–378. doi:10.1111/j.0956-7976.2004.00687.x

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology, 24*, 645–662. doi:10.1016/j.appdev.2003.09.002

Hall, J. A., & Carter, J. D. (1999). Gender-stereotype accuracy as an individual difference. *Journal of Personality and Social Psychology, 77*, 350–350-359. doi:10.1037/0022-3514.77.2.350

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11*, 365–371. doi:10.1111/1467-9280.00272

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16*, 175–179. doi:10.1111/j.0956-7976.2005.00799.x

Kirnan, J. P., Alfieri, J. A., Bragger, J. D., & Harris, R. S. (2009). An investigation of stereotype threat in employment tests. *Journal of Applied Social Psychology, 39*, 359–388. doi:10.1111/j.1559-1816.2008.00442.x

Lanning v. Southeastern Pennsylvania Transportation Authority, 181 F. 3d 478 (3d Cir. 1999).

Lanning v. Southeastern Pennsylvania Transportation Authority, 308 F. 3d 286 (3d Cir. 2002).

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72*, 275–275-287. doi:10.1037/0022-3514.72.2.275

Logel, C., Iserman, E. C., Davies, P. G., Quinn, D. M., & Spencer, S. J. (2009). The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology, 45*, 299–312. doi:10.1016/j.jesp.2008.07.016

Martinot, D., & Désert, M. (2007). Awareness of a gender stereotype, personal beliefs and self-perceptions regarding math ability: When boys do not surpass girls. *Social Psychology of Education, 10*, 455–471. doi:10.1007/s11218-007-9028-9

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin, 28*, 1183–1193. doi:10.1177/01461672022812004

Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: The interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology, 88*, 432–446. doi:10.1037/0022-3514.88.3.432

Mayer, D. M., & Hanges, P. J. (2003). Understanding the stereotype threat effect with "culture-free" tests: An examination of its mediators and measurement. *Human Performance, 16*, 207–230. doi:10.1207/S15327043HUP1603_3

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology, 38*, 115–142. doi:10.1037/0012-1649.38.1.115

McFarland, L. A., Lev-Arey, D. M., & Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance, 16*, 181–205. doi:10.1207/S15327043HUP1603_2

McFarland, L. A., Ryan, A. M., & Kriska, S. D. (2003). Impression management use and effectiveness across assessment methods. *Journal of Management, 29*, 641–661.

McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology, 39*, 83–90. doi:10.1016/S0022-1031(02)00513-9

Meyer, B., & Melchers, K. G. (2010). *Stereotype threat in personnel selection: Evidence from a simulated selection setting and from high-stakes employment testing.* Unpublished manuscript.

Nguyen, H.-H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16*, 261–293. doi:10.1207/S15327043HUP1603_5

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*, 1314–1334. doi:10.1037/a0012702

Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16*, 231–259. doi:10.1207/S15327043HUP1603_4

Reeve, C. L., & Hakel, M. D. (2000). Toward an understanding of adult intellectual development: Investigating within-individual convergence of interest and knowledge profiles. *Journal of Applied Psychology, 85*, 897–908. doi:10.1037/0021-9010.85.6.897

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science, 17*, 31–35. doi:10.1111/j.1467-8721.2008.00543.x

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565–606. doi:10.1177/014920630002600308

Rydell, R. J., McConnell, A. R., & Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology, 96*, 949–966. doi:10.1037/a0014846

Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance, 16*, 295–309. doi:10.1207/S15327043HUP1603_6

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist, 59*, 7–13. doi:10.1037/0003-066X.59.1.7

Sackett, P. R., & Ryan, A. M. (2011). Concerns about generalizing stereotype threat research findings to operational high-stakes testing settings. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 249–263). New York, NY: Oxford University Press.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318. doi:10.1037/0003-066X.56.4.302

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38*, 194–201. doi:10.1006/jesp.2001.1500

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115*, 336–356. doi:10.1037/0033-295X.115.2.336

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of employment selection procedures.* Bowling Green, OH: Author.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4–28. doi:10.1006/jesp.1998.1373

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613–629. doi:10.1037/0003-066X.52.6.613

Steele, C. M. (2010). *Whistling Vivaldi and other clues to how stereotypes affect us.* New York, NY: Norton.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811. doi:10.1037/0022-3514.69.5.797

Steele, C. M., & Davis, P. G. (2003). Stereotype threat and employment testing: A commentary. *Human Performance, 16*, 311–326. doi:10.1207/S15327043HUP1603_7

Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology, 77*, 1213–1227. doi:10.1037/0022-3514.77.6.1213

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34*, 665–693. doi:10.1111/j.1559-1816.2004.tb02564.x

Walker, M. E., & Bridgeman, B. (2008). *Stereotype threat spillover and SAT scores* (Research Report No. 2008–2.). New York, NY: College Board.

Walters, A. M., Lee, S., & Trapani, C. (2004). *Stereotype threat, the test-center environment, and performance on the GRE General Test* (GRE Board Research Report No. 01-03R.) Princeton, NJ: Educational Testing Service.

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39*, 456–467. doi:10.1016/S0022-1031(03)00019-2

CHAPTER 37

JOB SATISFACTION AND OTHER JOB ATTITUDES

Reeshad S. Dalal and Marcus Credé

An *attitude* is a favorable or unfavorable evaluation of a particular entity or object (Eagly & Chaiken, 1993). In this chapter, we are interested in the way in which employees evaluate their jobs. Researchers have suggested that there are several job-related attitudes. Of these, by far the most important is job satisfaction, which has been investigated in nearly 29,000 research studies (according to a PsycINFO search on January 16, 2011)—more than twice as often as all the other job attitudes put together. Indeed, Roznowski and Hulin (1992) have argued persuasively that, once an individual joins an organization, the most informative information an organizational psychologist or manager can possess about this individual is his or her level of job satisfaction. Furthermore, as we discuss subsequently, distinguishing the various job attitudes from each other empirically is often difficult. We therefore emphasize job satisfaction and discuss the other job attitudes only briefly.

Job satisfaction is defined as a multidimensional favorable or unfavorable response to the job situation (Judge, Hulin, & Dalal, 2012). The prevalent approach to the measurement of job satisfaction is based on three broad theoretical foundations: (a) Job satisfaction is organized hierarchically, with satisfaction with specific facets (aspects) of the job underpinning a single broad, general Job Satisfaction factor; (b) job satisfaction, like other attitudes, has a cognitive component and an affective (emotional) component; and (c) job satisfaction, and particularly its affective component, exhibits meaningful change over time within a given person. In this chapter, we discuss these three foundations, describe well-known ways of measuring cognitive and affective reactions to the job, describe best practices for attitude measurement and, finally, discuss some important areas for future measurement-related research.

COGNITIVE VERSUS AFFECTIVE COMPONENTS OF JOB SATISFACTION

The classical definition of attitudes (e.g., Thurstone, 1928) includes cognitive, affective, and behavioral components. However, we—as have many others before us (Chaiken & Stangor, 1987; Judge et al., 2012; Wyer, 1974)—maintain that the inclusion of behavior (i.e., overt action) in the very definition of attitudes is inimical to the study of attitude–behavior relationships. Behavior should be conceptualized as a correlate of attitudes (e.g., a consequence or a cause of attitudes), not as a component of attitudes. In other words, observations of behavior inform one about the expressed behavior, but they do not directly inform one about the relevant attitudes (which need to be measured independently). Therefore, in this chapter we define *job satisfaction* as a set of cognitive and affective responses to the job situation. This definition is consistent with the typical conceptualization of job attitudes as important predictors of job-related behavior—in

The authors contributed equally; the order of names in the byline is arbitrary.

DOI: 10.1037/14047-037
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

particular, of performance on the job and behavior related to quitting the job (Dalal, 2005; Griffeth, Hom, & Gaertner, 2000; Judge, Thoresen, Bono, & Patton, 2001).

Although job satisfaction ostensibly consists of both cognitive and affective components, both of which should exist in measures of job satisfaction, in practice most measures of job satisfaction focus primarily on the cognitive component (Judge et al., 2012; H. M. Weiss, 2002). The affective component has, in other words, historically received short shrift vis-à-vis job satisfaction. This has at least two major measurement-related consequences: (a) what, in terms of content, is measured and (b) the level of analysis at which this content is measured.

Content of Measurement

We discuss the content of measurement separately for cognitive and affective reactions to the job situation.

Cognitive reactions. Many of the well-known measures of job satisfaction focus primarily on employees' cognitive descriptions, evaluations of various facets of the job, or both. Numerous facets of the job can, and have, been studied, but most of these cognitive measures of job satisfaction include facets such as the supervisor, coworkers, amount of pay and benefits, opportunities for promotion, and the (nature of the) work itself.

Studies have often simply averaged (or summed) facet satisfaction scores in an attempt to assess overall job satisfaction. This practice is, however, undesirable from a conceptual standpoint because it involves the following assumptions, none of which is likely to be defensible: (a) the assumption that all facets relevant to every employee's job are measured and that no facet irrelevant to any employee's job is measured (i.e., that there are no errors of omission and commission, respectively), (b) the assumption that the various facets should be weighted equally in determining overall job satisfaction, and (c) the assumption that facets combine in a linear, additive fashion in determining overall job satisfaction (Balzer et al., 2000; Ironson, Smith, Brannick, Gibson, & Paul, 1989; Scarpello & Campbell, 1983). Instead, overall job satisfaction is best assessed via global measures that ask employees to respond vis-à-vis their jobs as a whole (rather than vis-à-vis individual facets of their jobs).

Under what circumstances should global measures be used in lieu of facet measures? Research in social psychology (e.g., Ajzen, 2005) and industrial and organizational psychology (e.g., Lavelle, Rupp, & Brockner, 2007) has suggested that attitudes predict behavior best when the attitude and behavior are at the same level of generality (i.e., granularity) and when they are directed toward the same object (i.e., target). Thus, for example, employees' deviant behavior directed toward their supervisor should be better predicted by their satisfaction with the supervisor than by their overall (i.e., global) job satisfaction. In contrast, employees' overall deviant behavior should be better predicted by their overall job satisfaction than by their satisfaction with the supervisor. The point here is simply that neither a global nor a facet measure of satisfaction is inherently better than the other: The utility of both global and facet measures depends on the specific behavior being predicted, and both are important for a thorough understanding of employees' responses to the job situation.

Affective reactions. Affective reactions are typically studied as moods, discrete emotions, or both. Moods are thought to be less intense but to persist for a longer duration than emotions; moreover, moods, unlike emotions, are not directed at specific people or objects (H. M. Weiss & Cropanzano, 1996).

The structure of mood is generally believed to reduce to two dimensions. However, which two dimensions are implicated is a relatively contentious issue. According to one school of thought (Feldman Barrett & Russell, 1998), mood consists of the dimensions of hedonic tone (pleasantness–unpleasantness) and activation (intensity). These dimensions are conceptualized as bipolar; therefore, the opposite of a pleasant mood is an unpleasant mood, and the opposite of an intense mood is a mild mood. According to the second school of thought (Watson & Clark, 1999), mood consists of the dimensions of positive affect and negative affect. These dimensions are conceptualized as unipolar; therefore, the opposite of a positive mood is not a

negative mood but rather the absence of a positive mood, and the opposite of a negative mood is not a positive mood but rather the absence of a negative mood. An extensive discussion of the merits and demerits of each of these structures is well beyond the scope of this chapter. However, we make two observations in passing. First, some conceptual awkwardness notwithstanding (see H. M. Weiss & Cropanzano, 1996), the latter structure appears to have become the dominant one in industrial and organizational psychology. Second, there is some reason to believe that the difference between the two structures is more apparent than real (for details, see Tellegen, Watson, & Clark, 1999; H. M. Weiss & Cropanzano, 1996).

The structure of mood can be contrasted with the structure of discrete emotions. In the case of emotions, as noted by H. M. Weiss and Cropanzano (1996), "The problem . . . is not a lack of structure but, instead, a surfeit of perspectives, points-of-view, and theoretical models" (pp. 19–20). Numerous attempts have been made to identify basic or primary emotions, but findings have differed somewhat across studies, in part because researchers have adopted different philosophical perspectives (e.g., evolutionary, physiological, and semantic). A review of extant taxonomies is provided by H. M. Weiss and Cropanzano.

It is important to note that affective reactions are not confined to the job situation. For example, people are likely to have affective reactions to their family, news headlines, medical procedures they are undergoing, and so forth. However, affective reactions to the job can be viewed as a complement to cognitive reactions to the job—and hence as an important aspect of job satisfaction that is missed by traditional measures, which focus primarily on cognitive reactions.

We next discuss the level of analysis, insofar as it is relevant to the measurement of cognitive and affective reactions to the job situation.

Level of Analysis

Traditionally, cognitively oriented conceptualizations of job satisfaction have focused on the person level of analysis. Each person's job satisfaction is typically measured only once in a particular study, and the comparison is across people (e.g., "Is Jane's job satisfaction higher than Jill's job satisfaction?"). Exceptions do exist, of course. The most common alternative to the person level has been the unit level (e.g., Whitman, Van Rooy, & Viswesvaran, 2010). Here, the aggregated job cognitions of employees are compared across work units (e.g., "Are employees in the human resources department more satisfied than employees in the marketing department?") using appropriate theoretical models of how unit-level job satisfaction is composed of person-level job satisfaction (see Chan, 1998).

Affective reactions to the job, too, can be assessed at the person level. However, they have also often been assessed at the within-person level. At the within-person level, each person's affective reactions are typically measured on several different occasions during the workday, and the comparison is across time within the same person (e.g., "Is Jimmy's mood at work more negative now than it was 2 hours ago?"). Of course, the appropriate level of analysis depends on the research question. Affective reactions to specific workplace events (e.g., spilling coffee all over one's clothes, receiving unexpected praise from one's supervisor) should be studied at the within-person level. However, habitual patterns of affectivity (e.g., a general tendency to feel enthusiastic at work) should be studied at the person level.

Several measures of both cognitive and affective reactions to the job have already been developed. In the next several sections, we provide an overview of the best known of these measures. First, we discuss well-known measures of job satisfaction. Unfortunately, most of these measures emphasize cognitive reactions at the expense of affective reactions. Second, we discuss well-known measures of affective reactions. Third, we discuss well-known measures of job attitudes other than job satisfaction. After that, we evaluate the strengths and limitations of all of these measures in the course of describing best practices in attitude measurement as well as areas in which future research is needed. Throughout these sections, we emphasize self-report measures (i.e., employees answering questions regarding their own job attitudes) because self-reports have been the predominant approach to job attitude measurement. Toward the end of this chapter, however, we briefly discuss alternatives to self-report measures.

WELL-KNOWN MEASURES OF JOB SATISFACTION

In this section, we primarily discuss four well-known measures of job satisfaction: the Job Descriptive Index (JDI), the Job in General (JIG) scale, the Minnesota Satisfaction Questionnaire (MSQ), and the Faces scale. The first three primarily measure cognitive reactions to the job, whereas the fourth measures a combination of affective and cognitive reactions (Brief & Roberson, 1989; Fisher, 2000; Schleicher, Watt, & Greguras, 2004).

Job Descriptive Index and Job in General Scales

The JDI (Smith, Kendall, & Hulin, 1969) is perhaps the most widely used and widely studied facet measure of job satisfaction (Balzer et al., 2000; Judge et al., 2001). The JDI measures employee satisfaction with five facets of the job: the work itself, supervision, coworkers, pay, and opportunities for promotion. These five facets have the advantage of being applicable to employees at virtually all levels of the organizational hierarchy (Balzer et al., 2000). Of the five facets, research has suggested that satisfaction with the work itself is by far the most important vis-à-vis overall (global) job satisfaction (Ironson et al., 1989).

Satisfaction with the work itself, supervision, and coworkers are each measured with 18 items. Satisfaction with pay and promotions are each measured with nine items. In total, therefore, the JDI consists of 72 items, each of which consists of a single word or a short phrase describing the attitude object in question. Participants respond to each item using the following response options: "Yes" (if the item describes the job facet), "No" (if the item does not describe the job facet), and "?" (if the participant cannot decide whether the item describes the job facet). On primarily empirical grounds, the "?" option is typically scored as being twice as close to the option indicating dissatisfaction as to the option indicating satisfaction (Hanisch, 1992). A shorter version of the JDI, with five items per facet, has been developed (Stanton et al., 2001).

As noted in the section Content of Measurement earlier in this chapter, facet satisfaction measures (such as the JDI) provide different information than global satisfaction measures. Consequently, although the JDI facet scores are often averaged to yield a global job satisfaction score, a much better approach to measuring global job satisfaction involves the use of the JIG scale (Ironson et al., 1989), which typically accompanies the JDI. The JIG has 18 items, and its response options and instructions parallel those of the JDI. A shortened version of the JIG contains only eight items (S. R. Russell et al., 2004). Norms for specific categories (e.g., job level and organization type) exist for both the JDI and the JIG (Balzer et al., 2000).

In recent years, researchers have also developed JDI-like measures of additional facets of the job—for example, satisfaction with job security (Probst, 2003) and satisfaction with management above the level of the immediate supervisor (Dalal, Bashshur, & Credé, 2011)—that, although not covered by the JDI, have been demonstrated to be important to employees (perhaps because of the changing nature of employer–employee relationships since the publication of the JDI). The new measures developed by Probst (2003) and by Dalal et al. (2011) have been modeled on the JDI in a structural sense (in terms of response options, instructions, and type of items) but are nonetheless empirically distinguishable from the JDI facets.

More details regarding the JDI and JIG—including sample items, information about abridged versions of the scales, a brief history of the scales, a way to access archival datasets using these scales, and so forth—can be obtained at the following website: http://www.bgsu.edu/departments/psych/io/jdi.

Minnesota Satisfaction Questionnaire

Another well-known facet measure of job satisfaction is the MSQ (D. J. Weiss, Dawis, England, & Lofquist, 1967). Participants respond to items by indicating how satisfied they are with several aspects of the job environment. Two sets of response options have been used. The first set of response options consists of the following five options: *very dissatisfied, dissatisfied, neither satisfied nor dissatisfied, satisfied,* and *very satisfied.* However, job satisfaction scores using these response options demonstrated a negative skew, such that very few employees claimed to be very dissatisfied. Consequently, the following set of

response options was formulated: *not satisfied, somewhat satisfied, satisfied, very satisfied,* and *extremely satisfied.*

The MSQ has two forms: a long form and a short form (D. J. Weiss et al., 1967). The long form includes 20 facets, each measured with five items. Every facet measured by the JDI is also measured by the MSQ using either a single facet or a combination of facets. In addition, the MSQ covers other facets, such as "company policies and practices," "authority," and "social service."

The short form of the MSQ includes 20 items, with one item being selected from each of the 20 facets on the long form (from each facet, the item whose scores correlate most highly with scores on the facet as a whole is chosen). Three scores are extracted from the short form: extrinsic (environmental) satisfaction, intrinsic (direct experience) satisfaction, and general satisfaction (calculated using all 20 items). More details regarding the MSQ—including the administration manual, with a complete list of items composing the long and short forms—can be obtained at the following website: http://www.psych.umn.edu/psylabs/vpr/msqinf.htm.

Faces Scale

The Faces scale (Kunin, 1955) is a single-item measure of global job satisfaction. The scale presents a series of faces that vary from extremely unhappy to extremely happy and asks the respondents to indicate which face best represents how they feel about their job in general. The full scale contains 11 faces, but typically only five or seven are used (Dunham & Herman, 1975; Kunin, 1955). Additional measurement-related details—such as whether to use abstract faces or human faces and, if the latter, whether the gender of the faces makes a difference—are discussed in Kunin (1955) and Dunham and Herman (1975).

Other Measures of Job Satisfaction

Numerous other measures of job satisfaction have also been developed, most of which also focus primarily on cognitive rather than affective reactions. Two of the better known measures are the Brayfield–Rothe measure (Brayfield & Rothe, 1951) and the Job Satisfaction Survey (Spector, 1985; for more details, including a list of items, see the following website: http://shell.cas.usf.edu/~pspector/scales/jsspag.html). These and many other measures of job satisfaction are reviewed in Cook, Hepworth, Wall, and Warr (1981) and Fields (2002), who also provide complete lists of items and information regarding reliability and validity. In addition, measures have been developed to assess the job satisfaction of employees in specific occupations such as nursing (e.g., Mueller & McCloskey, 1990) and social work (Shapiro, Burkey, Dorman, & Welker, 1997).

WELL-KNOWN MEASURES OF AFFECTIVE REACTIONS

In this section, we describe existing measures of affective reactions. At the outset, however, we remind the reader that a person can have affective reactions not just to the job situation but to several other domains as well (e.g., the family situation). Therefore, measures of affective reactions have typically not been designed solely with the job as a frame of reference. However, industrial and organizational psychologists often use these measures with the job as an imposed frame of reference because of the paucity of existing affectively oriented measures of job satisfaction.

Positive and Negative Affect Schedule—Expanded Form

Perhaps the best-known measure of affect is the Positive and Negative Affect Schedule—Expanded Form (PANAS–X; Watson & Clark, 1999). The PANAS–X consists of 60 items that measure mood and specific emotions at different levels of abstraction. At the higher level of abstraction, two unipolar dimensions of mood—namely, positive affect and negative affect—are measured. The dimensions of positive affect and negative affect are often believed to be unrelated to each other, but empirical evidence has suggested that the correlation between them is roughly −.40 (J. A. Russell & Carroll, 1999; Tellegen et al., 1999). However, the correlation is artificially lowered when positive and negative affect are measured using the PANAS–X (rather than other questionnaires), because items in the PANAS–X were deliberately chosen to be pure markers of one dimension or the other (see Watson & Clark, 1999).

At the lower level of abstraction, the PANAS–X measures four basic negative emotions, three basic positive emotions, and four other affective states. The four basic negative emotions are strongly positively interrelated and, consequently, compose the higher order factor of Negative Affect. The four basic positive emotions are likewise strongly positively interrelated and, consequently, compose the higher order factor of Positive Affect. The other affective states cannot readily be mapped on to either Positive Affect or Negative Affect.

The items in the PANAS–X can be administered using no fewer than eight possible sets of instructions regarding the time interval across which respondents are asked to report how they feel (or have felt), ranging from "right now (that is, at the present moment)" to "in general, that is, on the average" (Watson & Clark, 1999, p. 3). In organizational research, the instructions are frequently augmented to emphasize that respondents should respond with regard to how they feel on the job. Thus, researchers can choose the instructions that are appropriate for their research questions and desired levels of analysis.

The administration manual for the PANAS–X, which includes a complete list of items, can be found at the following website: http://www.psychology.uiowa.edu/Faculty/Clark/PANAS-X.pdf.

Other Measures of Affective Reactions

Numerous other measures of affective reactions exist. Feldman Barrett and Russell (1998) provided several measures of mood, including those designed to measure the two bipolar dimensions of hedonic tone and activation (rather than positive and negative affect). A measure of specific emotions is provided by Shaver, Schwartz, Kirson, and O'Connor (1987). These measures, like the PANAS–X, are not inherently job specific; however, they too can be modified for this purpose (thereby addressing the paucity of measures of affective reactions to the job).

JOB ATTITUDES OTHER THAN JOB SATISFACTION

As mentioned previously, job satisfaction is not the only job attitude studied by organizational researchers (although it is by far the most heavily studied one). Other constructs that share some of the evaluative and affective characteristics of job attitudes but that differ with respect to the object (target) at which the attitudes are directed include attitudes for which the target is (a) the job, (b) the organization, and (c) the work being performed. Job involvement (Paullay, Alliger, & Stone-Romero, 1994) and employee engagement (Macey & Schneider, 2008) fall into the first of these categories. Organizational commitment (Mowday, Steers, & Porter, 1979), perceived organizational support (Rhoades & Eisenberger, 2002), and perceptions of organizational justice (Colquitt, 2001) fall into the second category. Work centrality (Paullay et al., 1994) falls into the last category. Although distinct literatures have developed for each of these attitudes, some debate has occurred as to whether each attitude is truly conceptually distinct from other attitudes directed toward the same object or even different objects (e.g., Little & Little, 2006). Measures of these attitudes often contain items very similar to those included in measures of other attitudes (Newman & Harrison, 2008). Moreover, the empirical relationships among these job attitudes are sufficiently strong—for example, $\rho = .73$ for the relationship between perceived organizational support and affective organizational commitment (Rhoades & Eisenberger, 2002) and $\rho = .91$ for the business unit–level relationship between employee engagement and job satisfaction (Harter, Schmidt, & Hayes, 2002)—that some doubt exists as to whether the newer and currently more fashionable job attitudes, such as employee engagement, add value beyond the traditionally studied job attitudes, such as job satisfaction (Little & Little, 2006; see also Dunnette, 1966). This has led some authors (e.g., Credé, 2005; Harrison, Newman, & Roth, 2006; Le, Schmidt, Harter, & Lauver, 2010; Newman & Harrison, 2008; Newman, Joseph, & Hulin, 2010) to propose and empirically confirm a hierarchical structure to job attitudes, with a general factor explaining the high covariation among more specific job attitudes. Because not all attitudinal constructs have yet been included in tests of such a hierarchical model, future work should continue to examine the degree to which a single general factor

can account for the covariation among the various attitudinal constructs examined in the organizational literature. Despite the possible existence of a single overall factor, we briefly review two of the most notable of these additional job attitudes, describe how they are typically measured in the organizational sciences, and describe the relationships among them.

Organizational Commitment

Organizational commitment refers to employees' attachment to the organization and identification with its goals. Originally conceptualized and measured as a unidimensional construct (see, e.g., the Organizational Commitment Questionnaire; Mowday et al., 1979), organizational commitment is now widely considered to consist of three components (Meyer & Allen, 1991): affective commitment (emotional attachment to the organization), normative commitment (perceived obligation to the organization), and continuance commitment (perceived costs associated with leaving the organization). This conceptualization of organizational commitment, in other words, captures not only cognitive reactions to the organization (i.e., continuance and normative commitment) but also affective reactions (i.e., affective commitment). Indeed, reviews of existing research have suggested that the affective component has the strongest relationship with work behavior (Meyer, Stanley, Herscovitch, & Topolnytsky, 2002). Widely used 24-item and 18-item self-report measures of affective, normative, and continuance commitment are described by Allen and Meyer (1990) and Meyer and Allen (1997), respectively, although the theoretical basis of the three-component model of commitment and the item content of the scales has recently been criticized (e.g., Solinger, van Olffen, & Roe, 2008).

In addition, a major limitation of the popular measures of organizational commitment is that they are contaminated with items pertaining to behavioral intentions—specifically, intentions to quit the job (Bozeman & Perrewé, 2001). Consequently, when organizational commitment is used as a predictor of intentions to quit, the obtained relationship may be spuriously high (because intentions to quit are inadvertently being predicted by themselves). This limitation underscores our previous assertion that measures of attitudes should contain only cognitive and affective components; behavior (or behavioral intention) should be treated as a correlate, not a component, of an attitude.

Employee Engagement

Employee engagement is perhaps the most recent job attitude to be studied by researchers. Consensus regarding the precise nature of employee engagement is still developing (see Macey & Schneider, 2008, for review), but commonly discussed characteristics include feelings of enthusiasm, energy, vigor, dedication, and absorption with regard to work tasks and roles. Several measures of employee engagement exist, including the 17-item Utrecht Work Engagement Scale (Schaufeli, Salanova, Gonzàlez-Romà, & Bakker, 2002) and the 11-item Job Engagement Scale (Saks, 2006). It is noteworthy that measures of employee engagement often include a focus on affective reactions in addition to cognitive reactions. For example, several of the items in the Utrecht Work Engagement Scale are essentially positive affect items from the aforementioned PANAS–X adapted for use vis-à-vis the job situation (Newman & Harrison, 2008).

As with organizational commitment, however, a major limitation of some measures of employee engagement (e.g., the May, Gilson, & Harter, 2004, measure) is that they are contaminated with behavioral items—in this case, with items involving positive behavior such as organizational citizenship behavior (Dalal, Baysinger, Brummel, & LeBreton, in press; Dalal, Brummel, Wee, & Thomas, 2008). Consequently, when employee engagement is used as a predictor of organizational citizenship behavior, the obtained relationship may be spuriously high (because citizenship behavior is inadvertently being predicted by itself).

We divide the rest of this chapter into two broad sections. The first section is our attempt to delineate best practices for attitude measurement on the basis of lessons learned from industrial and organizational psychology and other fields (e.g., social psychology). As a part of this section, we describe the strengths and limitations of the attitude measures described previously. The second section is intended as an overview of what we consider to be emerging

trends and important areas of future inquiry in the measurement of job attitudes.

BEST PRACTICES FOR ATTITUDE MEASUREMENT

When developing a new measure of an attitude or evaluating an existing measure, several factors should be considered. In the sections that follow, we discuss some of the more important factors.

Inclusion or Exclusion of Reverse-Scored Items

The debate regarding the merits of including reverse-scored items has a substantial history. On one hand, reverse-scored items help reduce the impact of yea-saying or nay-saying on scale-level scores and may facilitate the detection of random or careless response patterns via an examination of the consistency of responses to negatively and positively worded items. On the other hand, in exploratory factor analyses of item-level attitude data, negatively worded items have frequently been observed to combine to form a separate negative-item factor that is widely considered to be artifactual (e.g., Harvey, Billings, & Nilan, 1985; Idaszak & Drasgow, 1987). Simulations (e.g., Schmitt & Stults, 1985) have shown that distinct negative-item factors can emerge when as few as 10% of the sample respond carelessly to attitude inventories, and other research (e.g., Green, Armenakis, Marbert, & Bedeian, 1979) has suggested that negative-item factors are particularly likely to emerge among samples with low education levels because of the greater difficulty associated with responding to negatively worded items in the intended direction. Given the extensive debate regarding the desirability of including negatively worded items, it is perhaps unsurprising that, whereas some of the measures we previously reviewed (e.g., the JDI) contain such items, other measures (e.g., the PANAS–X) do not.

Recent work in the attitude domain (Credé, Chernyshenko, Bagraim, & Sully, 2009), however, has suggested that negatively worded items may substantially increase the ability of job attitude measures to predict work behavior (i.e., criterion-related validity), as has evidence relating to Cacioppo and Berntson's (1994) bivariate evaluation plane (discussed in more detail later). We therefore advocate for the continued inclusion of negatively worded items in attitude scales, albeit with some important caveats.

One caveat is that negatively worded items should be characterized by words that denote negative attitudes rather than simple negations (*not, no*) of positive attitudes. For example, when reversing the item "I love my job," the reversed form "I hate my job" would be preferred to the reversed form "I do not love my job" for two reasons. First, employees who hate their jobs and those who feel neutral toward their jobs could legitimately agree with the reversed version "I do not love my job." In other words, *not loving* is less likely than *hating* to be interpreted as the mirror image of *loving*. Second, when responding to the reversed version "I do not love my job," employees who do, in fact, love their jobs would be compelled to process a double negative (i.e., disagreeing that they do not love their jobs), which is cognitively taxing.

The other caveat is that researchers should ensure that factors composed primarily of negative items are not the result of careless or random responding. One way to do this is via the inclusion of validity scales, but these scales are frequently ineffective (for additional information on validity scales, refer to Volume 2, Chapter 11, this handbook). Instead, researchers should model methodological "wording-direction" factors in addition to substantive attitude factors (see, e.g., Kelloway & Barling, 1990). We also recommend that researchers examine whether negative item factors exhibit unique relationships with other constructs.

Reading Level

Job attitude researchers frequently work with populations characterized by low reading ability, poor education, or both. Researchers should therefore favor using measures explicitly designed to accommodate respondents with a wide variety of reading abilities and education levels. The JDI is considered to be a model of item readability (E. F. Stone, Stone, & Gueutal, 1990).

Number of Items

Job satisfaction measures should contain enough items to ensure high levels of internal consistency

(a form of reliability) and adequate content coverage (i.e., content validity). As a result, most widely used measures of job attitudes—other than Kunin's (1955) single-item Faces scale—are composed of multiple items per construct to be measured. Some evidence has suggested that the reliability of a scale is often too low when fewer than five items are used to measure a construct but, conversely, that reliability often does not increase appreciably with many more than five items (Hinkin, 1998). We would therefore suggest the use of five to seven items per facet. One caveat is that measures of extremely broad constructs are likely to have lower interitem correlations and, therefore, to require a somewhat larger number of items to achieve conventionally accepted levels of reliability. Nonetheless, it may be difficult to justify 18 items per facet, which is the number of items used in the JIG and some facets of the JDI. Unsurprisingly, shorter versions of these measures have been developed (as described previously).

Number and Nature of Response Options

A substantial literature has suggested that the number and nature of response options provided to respondents have nontrivial effects on the psychometric properties of the measures. Larger numbers of response options have been linked to higher predictive power, higher relationships with other measures of the same construct, higher tendency to "look like" the measure assesses what it purportedly assesses, and higher stability of scores across time (i.e., higher criterion-related validity, convergent validity, face validity, and test–retest reliability, respectively; Chang, 1994; Loken, Pirie, Virnig, Hinkle, & Salmon, 1987; Preston & Colman, 2000; Weng, 2004). Thus, in general, having too few response options is undesirable. However, some (albeit less conclusive) evidence has shown that having too many response options is also undesirable from the standpoint of reliability and validity (Krosnick, Judd, & Wittenbrink, 2005). Although definitive conclusions are difficult to offer, five to seven response options may be optimal (Krosnick et al., 2005). For almost all the attitude measures reviewed previously, the number of response options is between five and seven. The JDI, however, has only three response options.

Use and Scoring of Central Response Option

An issue related to the number of response options is whether a middle or central response option should be used. Concerns about the meaning and interpretation of the midpoint of a response option continuum are neither new nor specific to the domain of attitude measurement, particularly when that midpoint is labeled in a manner indicating neither agreement nor disagreement with the item stem. Selection of the midpoint for a typical attitude item can be the result of indifference (i.e., neither positive nor negative attitude), ambivalence (i.e., both positive and negative attitude), confusion about item meaning, a lack of a defined attitude, or an unwillingness to commit to a single response (DuBois & Burns, 1975; Shaw & Wright, 1967; M. H. Stone, 2004; see also Kulas, Stachowski, & Haynes, 2008). However, some evidence has shown that the presence of a middle response option increases the reliability and validity of ratings (Krosnick et al., 2005). Thus, overall, the use of a middle response option has both advantages and disadvantages.

Most attitudinal measures—including the ones we reviewed previously—include a middle response option. The JDI, however, provides a good illustration of one of the difficulties associated with the middle response option. The middle option of the JDI takes the form of a question mark ("?") that respondents are instructed to use when they are unsure whether they agree or disagree with an item. Using item response theory (discussed later in this chapter), Hanisch (1992) provided evidence to support the idea that individuals choosing the "?" are more likely to be dissatisfied than satisfied. In other words, the "?" in the JDI is not a neutral midpoint. Consequently, it is scored as being twice as close to the response option indicating dissatisfaction as to the response option indicating satisfaction. In contrast, the other attitude measures reviewed previously score the middle response option as equidistant between the two adjacent response options (although, to our knowledge, the optimal scoring of the middle response in these other measures has not been investigated empirically).

Faking and Social Desirability

Faking and socially desirable responding are not considered important threats to the validity of job attitude data because (a) such data are not typically used to make high-stakes decisions such as hiring, firing, or promotion and (b) unlike other types of attitudes (e.g., racial attitudes), item responses on job attitudes are not considered to be greatly discrepant from each other in terms of social desirability. Employees may exaggerate their dissatisfaction with elements of the job if they perceive that job redesign efforts are likely to be guided by job attitude data, but in general the relationship between social desirability and job attitudes appears to be relatively weak (Moorman & Podsakoff, 1992).

Random and Careless Responding

A more serious threat to the validity of job attitude data is the possibility that, as alluded to in the Inclusion or Exclusion of Reverse-Scored Items section, a nontrivial proportion of respondents to attitude measures may respond in a manner that is effectively random. In both organizational and university research settings, participation in research is often not entirely voluntary (e.g., managers may strongly encourage employee participation, college students typically receive credit for participation), and it is consequently likely that some respondents choose to respond in a careless manner (e.g., Beach, 1989). It can be shown that even a small proportion of randomly responding individuals can fundamentally alter the inferences drawn from data (Credé, 2010). Because random responses to individual items behave nonrandomly when aggregated to the level of the overall measure (courtesy of the central limit theorem), randomly responding individuals can exert substantial effects on the relationships observed between attitude measures and outcomes such as behavior. We therefore recommend that researchers be aware of this effect and screen participant responses for inconsistent response patterns using quantitative methods, such as those outlined by Karabatsos (2003).

AREAS OF EMERGING AND FUTURE RESEARCH

We turn now to areas that have not thus far received much attention in job attitude research. We nonetheless discuss these areas because we believe that future research in these areas is necessary for a fuller understanding of how job attitudes should be measured.

Measurement Invariance and Equivalence

Often, researchers are interested in comparing job satisfaction scores across different groups (subpopulations) of individuals: female versus male employees, Generation X versus Generation Y employees, ophthalmologists versus optometrists versus opticians, employees who complete the job satisfaction survey via paper and pencil versus on the Internet, and so forth. Before making such comparisons, it is important to determine whether the construct of job satisfaction exhibits measurement invariance (or equivalence) across the groups being compared (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000)—in other words, whether being satisfied means the same thing to members of the groups being compared. If it does not, comparisons across groups are fraught with difficulty. For example, Hu, Kaplan, and Dalal (2010) assessed the invariance of the JDI across white-collar versus blue-collar employees. They found, for example, that directly comparing white-collar and blue-collar employees using the JDI Coworkers scale may be difficult because white-collar employees, but not blue-collar employees, make a distinction between their coworkers' likability and their work habits. As another example, Candell and Hulin (1986) found that the invariance of the JDI deteriorated more when groups differing in both language and culture were compared than when groups differing in either language or culture were compared.

We therefore recommend that future job satisfaction research routinely evaluate measurement invariance before comparing groups of employees. A lack of invariance limits the types of comparisons that can be made across groups.

Item Difficulty and Discrimination Issues

Item response theory research in the domain of intelligence has shown that items can be described by how well they differentiate between individuals possessing differing levels of the underlying psychological construct being assessed (e.g., Drasgow,

Chernyshenko, & Stark, 2010; see also Chapter 6, this volume). Attitude measurement is also likely to benefit from the use of item response theory analysis to identify items with the appropriate level of discrimination. Item response theory analyses have been conducted on the JDI (for examples, see Carter, Dalal, Lake, Lin, & Zickar, 2011; Hanisch, 1992; Roznowski, 1989) but have in general been underused in research on job attitudes.

Item response theory is likely to be especially useful in cases in which attitude–behavior relationships are nonlinear. If, for example, the likelihood of quitting one's job is low for both moderate and high levels of job satisfaction but high for low levels of job satisfaction, then researchers interested in predicting quitting behavior may be better served by assessing job satisfaction with items that distinguish between dissatisfied and moderately satisfied employees rather than with items that distinguish between moderately satisfied and highly satisfied employees.

Bivariate Evaluation Plane

The vast majority of job attitudes researchers view job attitudes as being bipolar in nature, such that an employee's attitude toward a job feature can be mapped onto a single dimension ranging from maximally positive to maximally negative. An alternate paradigm that has received some empirical support (Credé et al., 2009) is that job attitudes can be mapped onto what Cacioppo and Berntson (1994) termed the *bivariate evaluation plane,* whereby an individual's attitude is best captured by its position on a two-dimensional plane, the respective axes representing positive and negative attitudes (for a related perspective, see Herzberg, 1966). From this perspective, behavior is best understood via a joint consideration of positive and negative attitudes toward the attitude object, each making an independent contribution to the prediction of the behavior. The traditional bipolar attitude perspective equates the attitudes of an individual possessing low positive and low negative attitudes to those of an individual possessing high positive and high negative attitudes, whereas the bivariate perspective treats these two cases as meaningfully different. The bivariate perspective is consistent with the aforementioned conceptualization of affect as two unipolar dimensions of positive and negative affect (see our previous description of the PANAS–X).

Affective–Cognitive Consistency

Our earlier discussion of the theoretical nature of job attitudes highlighted their affective and cognitive components. In practice, the affective component, when not ignored outright, has typically been combined with the cognitive component to form a single aggregate attitude measure. There are, however, two reasons why this aggregation may result in a significant loss of information. First, the affective and cognitive components of an attitude may each provide added value (beyond the other component) in explaining important criteria. Thus, workplace behavior (e.g., ignoring the supervisor's instructions) may be influenced by both the affective component of the attitude (e.g., hating the supervisor) and the cognitive component of the attitude (e.g., thinking that the supervisor is incompetent). Second, the degree of congruence between the affective and cognitive components of an attitude may contain valuable information. High levels of affective–cognitive consistency have been linked to a greater likelihood of acting in accordance with attitudes (e.g., Kraus, 1995; Schleicher et al., 2004).

More research on affective–cognitive consistency per se in job attitudes is needed. In addition, when it has been studied, affective–cognitive consistency has generally been assessed by taking the absolute difference between the rank-order position of an individual's score on a measure of the affective component of the attitude and the rank-order position of his or her score on a measure of the cognitive component of the attitude. However, the use of difference scores leads to interpretational difficulties; thus, future research in this domain would ideally rely on more appropriate analytic approaches (see Edwards, 2002).

Attitude Importance

Affective–cognitive consistency has been suggested to be merely one component of a broader construct known as *attitude strength.* Strong attitudes are stable across time, resistant to change, and exert influence on both information processing and behavior

(Krosnick & Petty, 1995). Nine other components of attitude strength have been identified: importance, accessibility, extremity, intensity, certainty, interest in relevant information, knowledge, direct experience, and latitude of rejection and noncommitment (Krosnick, Boninger, Chuang, Berent, & Carnot, 1993). Of these other components, only attitude importance has received much attention by job attitude researchers.

According to Locke's (1976) value-percept model, the importance attached to a job facet (e.g., pay) determines the degree to which a discrepancy between the desired and the actual state of the job facet influences satisfaction with that facet. Although this specific role of importance in determining facet satisfaction has found some empirical support (e.g., McFarlin & Rice, 1992), research has suggested that the importance of facets does not influence the relationship between facet satisfaction and overall job satisfaction (e.g., Jackson & Corr, 2002; Rice, Gentile, & McFarlin, 1991), which may be because importance ratings have already been integrated into the level of facet satisfaction (i.e., employees are unlikely to be dissatisfied on facets that are unimportant to them). Future research should examine this possibility further, perhaps via verbal protocol analysis (Ericsson & Simon, 1993). Future research should also examine other components of attitude strength (e.g., attitude accessibility) in the context of job attitudes.

Alternatives to Self-Reported Cognition and Affect

In this chapter, we have focused on self-reported attitudes because self-report is by far the dominant approach to attitude measurement. Three alternatives to conventional self-report measures are observational measures, physiological measures, and implicit attitude measures. Observational measures, used primarily to assess affective reactions, encompass the analysis of facial expressions, whole-body movements, and written or oral narratives (Kaplan, Dalal, & Luchman, 2012). Physiological measures—such as blood pressure reactivity, pulse rate reactivity, and cortisol measurement—could also potentially be used to assess employees' cognitive and affective reactions to the job. Measures of implicit attitudes (attitudes not susceptible to conscious control or even awareness; e.g., Greenwald, McGhee, & Schwartz, 1998; see also Project Implicit, 2012) are particularly popular in the study of social attitudes, especially those characterized by significant social desirability issues (e.g., racial attitudes).

Each of these alternatives, however, has its own disadvantages. For example, for observational measures to be valid, all of the following requirements must be met: (a) The person's emotional state must translate into observable behavior (e.g., the wrinkling near the eyes that is characteristic of genuine smiles); (b) this behavior must, in fact, be observed; and (c) the observer must be able to accurately infer the person's emotional state from the observed behavior (Chan, 2009; Kaplan et al., 2012). A concern regarding physiological measures is that they are unlikely to be pure indicators of cognition and affect, making interpretation difficult (Kaplan et al., 2012). For example, blood pressure is influenced by numerous factors other than cognition and affect (e.g., level of activity, nutritional factors, drugs, disease, hormonal imbalances; Kaplan et al., 2012). Implicit measures have historically been plagued by measurement inadequacies and conceptual questions (e.g., Bosson, Swann, & Pennebaker, 2000).

Thus, none of these alternative approaches is a panacea. These alternatives could, nonetheless, provide valuable information when individuals are unwilling or unable to self-report accurately (although, as suggested previously, this may be less of a problem with job attitudes than with other attitudes). Perhaps more important, these alternatives could provide a deeper conceptual understanding of job attitudes (e.g., the interplay between conscious and nonconscious attitudes, the physiological correlates of psychological attitudes). We therefore endorse their further study.

CONCLUSIONS

The measurement of job satisfaction has benefited significantly not only from decades of research by job satisfaction researchers but also from the willingness of the field to learn from developments in the study of other types of attitudes (e.g., social

attitudes) and developments in the study of measurement per se. Similarly, researchers studying other types of attitudes could learn much from developments in the measurement of job attitudes (see Judge et al., 2012). We anticipate that this fruitful cross-pollination of insights and ideas will continue in the future. For example, we believe that the measurement of job satisfaction will benefit from a better integration of attitude strength findings from the social psychology domain, a greater exploration of alternatives to traditional self-report measures of job satisfaction (e.g., physiological correlates of job satisfaction), and the application of recent developments in item response theory that allow intermediate levels of job satisfaction to be assessed more accurately and rapidly. We would also expect to see job satisfaction researchers paying workplace affect and the within-person level of analysis the attention they deserve. In general, then, we believe that job satisfaction (and its measurement) is likely to remain a vibrant research area for the foreseeable future and that the existing wealth of knowledge positions job satisfaction as the gold standard among job attitudes—not just for researchers but also for practitioners interested in evidence-based solutions.

References

Ajzen, I. (2005). Laws of human behavior: Symmetry, compatibility, and attitude–behavior correspondence. In A. Beauducel, B. Biehl, M. Bosniak, W. Conrad, G. Schönberger, & D. Wagener (Eds.), *Multivariate research strategies* (pp. 3–19). Aachen, Germany: Shaker Verlag.

Allen, N. J., & Meyer, J. P. (1990). The measurement and antecedents of affective, continuance, and normative commitment to the organization. *Journal of Occupational Psychology, 63*, 1–18. doi:10.1111/j.2044-8325.1990.tb00506.x

Balzer, W. K., Kihm, J. A., Smith, P. C., Irwin, J. L., Bachiochi, P. D., Robie, C., . . . Parra, L. F. (2000). Users' manual for the Job Descriptive Index (JDI; 1997 version) and the Job in General scales. In J. M. Stanton & C. D. Crossley (Eds.), *Electronic resources for the JDI and JIG* (pp. 3–90). Bowling Green, OH: Bowling Green State University.

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied, 123*, 101–103.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631–643. doi:10.1037/0022-3514.79.4.631

Bozeman, D. P., & Perrewé, P. L. (2001). The effect of item content overlap on Organizational Commitment Questionnaire–turnover cognitions relationships. *Journal of Applied Psychology, 86*, 161–173. doi:10.1037/0021-9010.86.1.161

Brayfield, A. H., & Rothe, H. F. (1951). An index of job satisfaction. *Journal of Applied Psychology, 35*, 307–311. doi:10.1037/h0055617

Brief, A. P., & Roberson, L. (1989). Job attitude organization: An exploratory study. *Journal of Applied Social Psychology, 19*, 717–727. doi:10.1111/j.1559-1816.1989.tb01254.x

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*, 401–423. doi:10.1037/0033-2909.115.3.401

Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17*, 417–440. doi:10.1177/0022002186017004003

Carter, N. T., Dalal, D. K., Lake, C. J., Lin, B. C., & Zickar, M. J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the Job Descriptive Index. *Organizational Research Methods, 14*, 116–146. doi:10.1177/1094428110363309

Chaiken, S., & Stangor, C. (1987). Attitudes and attitude change. *Annual Review of Psychology, 38*, 575–630. doi:10.1146/annurev.ps.38.020187.003043

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*, 234–246. doi:10.1037/0021-9010.83.2.234

Chan, D. (2009). So why ask me? Are self-report data really that bad? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 309–336). New York, NY: Routledge.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*, 205–215. doi:10.1177/014662169401800302

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology, 86*, 386–400. doi:10.1037/0021-9010.86.3.386

Cook, J. D., Hepworth, S. J., Wall, T. D., & Warr, P. B. (1981). *The experience of work: A compendium*

and review of 249 measures and their use. London, England: Academic Press.

Credé, M. (2005). *Job attitudes: Tests of utility and position*. Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*, 596–612. doi:10.1177/0013164410366686

Credé, M., Chernyshenko, O. S., Bagraim, J., & Sully, M. (2009). Contextual performance and the job satisfaction–dissatisfaction distinction: Examining artifacts and utility. *Human Performance, 22*, 246–272. doi:10.1080/08959280902970427

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*, 1241–1255. doi:10.1037/0021-9010.90.6.1241

Dalal, R. S., Bashshur, M. R., & Credé, M. (2011). The forgotten facet: Employee satisfaction with management above the level of immediate supervision. *Applied Psychology, 60*, 183–209. doi:10.1111/j.1464-0597.2010.00431.x

Dalal, R. S., Baysinger, M., Brummel, B. J., & LeBreton, J. M. (in press). The relative importance of employee engagement, other job attitudes, and trait affect as predictors of overall employee job performance. *Journal of Applied Social Psychology*.

Dalal, R. S., Brummel, B. J., Wee, S., & Thomas, L. L. (2008). Defining employee engagement for productive research and practice. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 52–55. doi:10.1111/j.1754-9434.2007.00008.x

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 465–476. doi:10.1111/j.1754-9434.2010.01273.x

Dubois, B., & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement, 35*, 869–884. doi:10.1177/001316447503500414

Dunham, R. B., & Herman, J. B. (1975). Development of a female Faces scale for measuring job satisfaction. *Journal of Applied Psychology, 60*, 629–631. doi:10.1037/0021-9010.60.5.629

Dunnette, M. D. (1966). Fads, fashions, and folderol in psychology. *American Psychologist, 21*, 343–352. doi:10.1037/h0023535

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich.

Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Drasgow & N. W. Schmitt (Eds.), *Advances in measurement and data analysis* (pp. 350–400). San Francisco, CA: Jossey-Bass.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology, 74*, 967–984. doi:10.1037/0022-3514.74.4.967

Fields, D. L. (2002). *Taking the measure of work*. Thousand Oaks, CA: Sage.

Fisher, C. D. (2000). Mood and emotions while working: Missing pieces of job satisfaction. *Journal of Organizational Behavior, 21*, 185–202. doi:10.1002/(SICI)1099-1379(200003)21:2<185::AID-JOB34>3.0.CO;2-M

Green, S. B., Armenakis, A. A., Marbert, L. D., & Bedeian, A. (1979). An evaluation of the response format and scale structure of the Job Diagnostic Survey. *Human Relations, 32*, 181–188. doi:10.1177/001872677903200206

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464–1480. doi:10.1037/0022-3514.74.6.1464

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26*, 463–488. doi:10.1177/014920630002600305

Hanisch, K. A. (1992). The Job Descriptive Index revisited: Questions about the question mark. *Journal of Applied Psychology, 77*, 377–382. doi:10.1037/0021-9010.77.3.377

Harrison, D. A., Newman, D. A., & Roth, P. L. (2006). How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences. *Academy of Management Journal, 49*, 305–325. doi:10.5465/AMJ.2006.20786077

Harter, J. K., Schmidt, F. L., & Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. *Journal of Applied Psychology, 87*, 268–279. doi:10.1037/0021-9010.87.2.268

Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the Job Diagnostics Survey: Good news and bad news. *Journal of Applied Psychology, 70*, 461–468. doi:10.1037/0021-9010.70.3.461

Herzberg, F. I. (1966). *Work and nature of man*. New York, NY: Thomas Y. Crowell.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104–121. doi:10.1177/109442819800100106

Hu, X., Kaplan, S., & Dalal, R. S. (2010). An examination of blue- versus white-collar workers' conceptualizations of job satisfaction facets. *Journal of Vocational Behavior, 76*, 317–325. doi:10.1016/j.jvb.2009.10.014

Idaszak, J. R., & Drasgow, F. (1987). A revision of the Job Diagnostic Survey: Elimination of a measurement artifact. *Journal of Applied Psychology, 72*, 69–74. doi:10.1037/0021-9010.72.1.69

Ironson, G. H., Smith, P. C., Brannick, M. T., Gibson, W. M., & Paul, K. B. (1989). Construction of a Job in General scale: A comparison of global, composite, and specific measures. *Journal of Applied Psychology, 74*, 193–200. doi:10.1037/0021-9010.74.2.193

Jackson, C. J., & Corr, P. J. (2002). Global satisfaction and facet description: The moderating role of facet importance. *European Journal of Psychological Assessment, 18*, 1–8. doi:10.1027//1015-5759.18.1.1

Judge, T. A., Hulin, C. L., & Dalal, R. S. (2012). Job satisfaction and job affect. In S. W. J. Kozlowski (Ed.), *The Oxford handbook of industrial and organizational psychology* (pp. 496–525). New York, NY: Oxford University Press.

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, *127*, 376–407. doi:10.1037/0033-2909.127.3.376

Kaplan, S., Dalal, R. S., & Luchman, J. (2012). Measurement of emotions. In R. Sinclair, M. Wang, & L. Tetrick (Eds.), *Research methods in occupational health psychology: State of the art in measurement, design, and data analysis* (pp. 61–76). New York, NY: Routledge.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298. doi:10.1207/S15324818AME1604_2

Kelloway, E. K., & Barling, J. (1990). Item content versus item wording: Disentangling role conflict and role ambiguity. *Journal of Applied Psychology, 75*, 738–742. doi:10.1037/0021-9010.75.6.738

Kraus, S. J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin, 21*, 58–75. doi:10.1177/0146167295211007

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology, 65*, 1132–1151. doi:10.1037/0022-3514.65.6.1132

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 21–78). Mahwah, NJ: Erlbaum.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1–24). Hillsdale, NJ: Erlbaum.

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22*, 251–259. doi:10.1007/s10869-008-9064-2

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology, 8*, 65–77. doi:10.1111/j.1744-6570.1955.tb01189.x

Lavelle, J. J., Rupp, D. E., & Brockner, J. (2007). Taking a multifoci approach to the study of justice, social exchange, and citizenship behavior: The target similarity model. *Journal of Management, 33*, 841–866. doi:10.1177/0149206307307635

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes, 112*, 112–125. doi:10.1016/j.obhdp.2010.02.003

Little, B., & Little, P. (2006). Employee engagement: Conceptual issues. *Journal of Organizational Culture: Communications and Conflict, 10*, 111–120.

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297–1349). Chicago, IL: Rand McNally.

Loken, B., Pirie, P., Virnig, K. A., Hinkle, R. L., & Salmon, C. T. (1987). The use of 0±10 scales in telephone surveys. *Journal of the Market Research Society, 29*, 353–362.

Macey, W. H., & Schneider, B. (2008). The meaning of employee engagement. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 3–30. doi:10.1111/j.1754-9434.2007.0002.x

May, D., Gilson, R., & Harter, L. (2004). The psychological conditions of meaningfulness, safety, and availability and the engagement of the human spirit at work. *Journal of Occupational and Organizational Psychology, 77*, 11–37. doi:10.1348/096317904322915892

McFarlin, D. B., & Rice, R. W. (1992). The role of facet importance as a moderator in job satisfaction processes. *Journal of Organizational Behavior, 13*, 41–54. doi:10.1002/job.4030130105

Meyer, J. P., & Allen, N. J. (1991). A three-component conceptualization of organizational commitment: Some methodological considerations. *Human Resource Management Review, 1*, 61–89. doi:10.1016/1053-4822(91)90011-Z

Meyer, J. P., & Allen, N. J. (1997). *Commitment in the workplace.* Thousand Oaks, CA: Sage.

Meyer, J. P., Stanley, D. J., Herscovitch, L., & Topolnytsky, L. (2002). Affective, continuance, and normative commitment to the organization: A meta-analysis of antecedents, correlates, and consequences. *Journal of Vocational Behavior, 61*, 20–52. doi:10.1006/jvbe.2001.1842

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behavior research. *Journal of Occupational and Organizational Psychology, 65*, 131–149. doi:10.1111/j.2044-8325.1992.tb00490.x

Mowday, R., Steers, R., & Porter, L. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior, 14*, 224–247. doi:10.1016/0001-8791(79)90072-1

Mueller, C. W., & McCloskey, J. C. (1990). Nurses' job satisfaction: A proposed measure. *Nursing Research, 39*, 113–117.

Newman, D. A., & Harrison, D. A. (2008). Been there, bottled that: Are state and behavioral work engagement new and useful construct "wines?" *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 31–35. doi:10.1111/j.1754-9434.2007.00003.x

Newman, D. A., Joseph, D. L., & Hulin, C. L. (2010). Job attitudes and employee engagement: Considering the attitude "A-factor." In S. Albrecht (Ed.), *The handbook of employee engagement: Perspectives, issues, research, and practice* (pp. 43–61). Cheltenham, England: Edward Elgar.

Paullay, I. M., Alliger, G. M., & Stone-Romero, E. F. (1994). Construct validation of two instruments designed to measure job involvement and work centrality. *Journal of Applied Psychology, 79*, 224–228. doi:10.1037/0021-9010.79.2.224

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1–15. doi:10.1016/S0001-6918(99)00050-5

Probst, T. M. (2003). Development and validation of the Job Security Index and the Job Security Satisfaction Scale: A classical test theory and IRT approach. *Journal of Occupational and Organizational Psychology, 76*, 451–467. doi:10.1348/096317903322591587

Project Implicit. (2012). *Implicit social cognition: Investigating the gap between intentions and action.* Retrieved from http://www.projectimplicit.net/index.html

Rhoades, L., & Eisenberger, R. (2002). Perceived organizational support: A review of the literature. *Journal of Applied Psychology, 87*, 698–714. doi:10.1037/0021-9010.87.4.698

Rice, R. W., Gentile, D. A., & McFarlin, D. B. (1991). Facet importance and job satisfaction. *Journal of Applied Psychology, 76*, 31–39. doi:10.1037/0021-9010.76.1.31

Roznowski, M. (1989). Examination of the measurement properties of the Job Descriptive Index with experimental items. *Journal of Applied Psychology, 74*, 805–814. doi:10.1037/0021-9010.74.5.805

Roznowski, M., & Hulin, C. (1992). The scientific merit of valid measures of general constructs with special reference to job satisfaction and job withdrawal. In C. J. Cranny, P. C. Smith, & E. F. Stone (Eds.), *Job satisfaction* (pp. 123–163). New York, NY: Lexington.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125*, 3–30. doi:10.1037/0033-2909.125.1.3

Russell, S. R., Spitzmüller, C., Lin, L. F., Stanton, J. M., Smith, P. C., & Ironson, G. H. (2004). Shorter can also be better: The abridged Job in General Measure. *Educational and Psychological Measurement, 64*, 878–893. doi:10.1177/0013164404264841

Saks, A. M. (2006). Antecedents and consequences of employee engagement. *Journal of Managerial Psychology, 21*, 600–619. doi:10.1108/02683940610690169

Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology, 36*, 577–600. doi:10.1111/j.1744-6570.1983.tb02236.x

Schaufeli, W. B., Salanova, M., Gonzàlez-Romà, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies, 3*, 71–92. doi:10.1023/A:1015630930326

Schleicher, D. J., Watt, J. D., & Greguras, G. J. (2004). Reexamining the job satisfaction–performance relationship: The complexity of attitudes. *Journal of Applied Psychology, 89*, 165–177. doi:10.1037/0021-9010.89.1.165

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210–222. doi:10.1016/j.hrmr.2008.03.003

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*, 367–373. doi:10.1177/014662168500900405

Shapiro, J. P., Burkey, W. M., Dorman, R. L., & Welker, C. J. (1997). Job satisfaction and burnout in child abuse professionals: Measure development, factor analysis, and job characteristics. *Journal of Child Sexual Abuse, 5*, 21–38. doi:10.1300/J070v05n03_02

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*, 1061–1086.

Shaw, M. E., & Wright, J. M. (1967). *Scales for the measurement of attitudes.* New York, NY: McGraw-Hill.

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago, IL: Rand McNally.

Solinger, O. N., van Olffen, W., & Roe, R. A. (2008). Beyond the three-component model of organizational commitment. *Journal of Applied Psychology, 93*, 70–83. doi:10.1037/0021-9010.93.1.70

Spector, P. E. (1985). Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey. *American Journal of Community Psychology, 13*, 693–713. doi:10.1007/BF00929796

Stanton, J. M., Sinar, E. F., Balzer, W. K., Julian, A. L., Thoresen, P., Aziz, S., . . . Smith, P. C. (2001). Development of a compact measure of job satisfaction: The abridged Job Descriptive Index. *Educational and Psychological Measurement*, *61*, 1104–1122.

Stone, E. F., Stone, D. L., & Gueutal, H. G. (1990). Influence of cognitive ability on responses to questionnaire measures: Measurement precision and missing response problems. *Journal of Applied Psychology, 75*, 418–427. doi:10.1037/0021-9010.75.4.418

Stone, M. H. (2004). Substantive scale construction. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 201–225). Maple Grove, MN: JAM Press.

Tellegen, A., Watson, D., & Clark, L. A. (1999). Further support for a hierarchical model of affect: Reply to Green and Salovey. *Psychological Science, 10*, 307–309. doi:10.1111/1467-9280.00159

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554. doi:10.1086/214483

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70. doi:10.1177/109442810031002

Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the Positive and Negative Affect Schedule—Expanded Form.* Retrieved from http://www.psychology.uiowa.edu/Faculty/Watson/PANAS-X.pdf

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). *Manual of the Minnesota Satisfaction Questionnaire.* Minneapolis, MN: Work Adjustment Project.

Weiss, H. M. (2002). Deconstructing job satisfaction: Separating evaluations, beliefs and affective experiences. *Human Resource Management Review, 12*, 173–194. doi:10.1016/S1053-4822(02)00045-1

Weiss, H. M., & Cropanzano, R. (1996). Affective events theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. In B. M. Staw & L. L. Cummings (Eds.), *Research in Organizational Behavior: An annual series of analytical essays and critical reviews* (Vol. 18, pp. 1–74). New York, NY: Elsevier.

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*, 956–972. doi:10.1177/0013164404268674

Whitman, D. S., Van Rooy, D. L., & Viswesvaran, C. (2010). Satisfaction, citizenship behaviors, and performance in work units: A meta-analysis of collective construct relations. *Personnel Psychology, 63*, 41–81. doi:10.1111/j.1744-6570.2009.01162.x

Wyer, R. S. (1974). *Cognitive organization and change: An information processing approach.* Oxford, England: Erlbaum.

CHAPTER 38

LEGAL ISSUES IN INDUSTRIAL TESTING AND ASSESSMENT

Paul J. Hanges, Elizabeth D. Salmon, and Juliet R. Aiken

Organizations frequently use professionally designed tests and assessments to help inform their personnel decisions (e.g., selection, training, promotion). The consequences of not conducting employment practices or making decisions in a fashion consistent with legal standards have increased over the past 60 years. Unfortunately, new information about legal standards comes from numerous sources (e.g., case law, federal and state legislation, agency guidelines), and psychologists can be overwhelmed trying to make sense of this information. Moreover, psychologists benefit from having an understanding of how these legal standards influence, as well as occasionally disagree with, the professional standards of the discipline of industrial and organizational psychology (see McCauley, 2011). In this chapter, we hope to facilitate the reader's understanding of some basic legal issues surrounding the construction and use of tests and assessments in organizational settings.

The chapter begins with a brief review of the historical roots of modern equal employment opportunity (EEO) law. This background provides a necessary foundation for understanding subsequent legal decisions and EEO programs. Next, the chapter presents a concise review of the three documents—the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission [EEOC], Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice, 1978), the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 2003)—that affect how organizations use tests and assessment tools. In the next section of the chapter, the authors review the legal concept of adverse impact and discuss the impact of legal decisions and guidelines on test validation and implementation efforts. In the final section, the authors explain the shifting-burden-of-proof model, which is used to determine the outcome of EEO court cases, and present a few recommendations regarding using tests to improve EEO opportunity.

A BRIEF HISTORY OF MODERN EQUAL EMPLOYMENT OPPORTUNITY LAW

EEO law is a constantly evolving corpus arising from executive, legislative, and judicial actions. The earliest manifestation of EEO law had its roots in the Reconstruction era and legislation passed after the end of the U.S. Civil War (e.g., the Civil Rights Act of 1866).[1] However, scholarly discussion has generally identified the 1940s as the starting point of modern EEO law (Jones, 1977). In 1941, before the United States' entry into World War II, A. Phillip

Elizabeth D. Salmon and Juliet R. Aiken contributed equally to this chapter.

[1]It can also be argued that its roots can be traced back even further to the Bill of Rights. In *Washington v. Davis* (1976), the plaintiffs claimed that an employment procedure violated their Fifth Amendment right to due process.

DOI: 10.1037/14047-038
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)
Copyright © 2013 by the American Psychological Association. All rights reserved.

Randolph, an African American civil rights leader, proposed a march on Washington to protest racial discrimination in industries involved with the war effort (Jones, 1977). To prevent this protest, President Roosevelt issued Executive Order 8802 (1941), which encouraged "full participation in the national defense program by all citizens of the United States, regardless of race, creed, color, or national origin."[2] This order applied to employment by the federal government, all defense contracts, and vocational and training programs administered by federal agencies (Gutman, Koppes, & Vadonovich, 2011). It established the first EEO enforcement agency, the Fair Employment Practice Committee in the Office of Production Management. However, as was the case with several EEO enforcement entities, the committee lacked the staff and direct enforcement powers to enact the executive order (Jones, 1977).

In 1943, Roosevelt attempted to strengthen the Fair Employment Practice Committee with Executive Order 9346, which provided the committee with broader jurisdiction and more staff (Gutman et al., 2011; Jones, 1977). Nevertheless, the committee still lacked direct enforcement power and thus relied on negotiation, moral persuasion, and the pressure of public opinion to enforce its decisions (Jones, 1977). Despite the difficulties enforcing EEO law during this time, pressure for fair employment practices continued to mount. Indeed, starting in the 1940s, eight states and some cities developed their own fair employment agencies (Guion, 1998). However, these agencies' effectiveness varied (Guion, 1998; Jones, 1977). Thus, until the 1960s, the extent to which EEO existed in the United States varied on a state-by-state and sometimes even a city-by-city basis. Some states were as ineffective in changing business practices as the federal government had been, whereas others (e.g., the District of Columbia) were able to explicitly prohibit particular forms of discrimination within their boundaries (Guion, 1998).

Not until President Kennedy's 1961 Executive Order 10925 were important steps again taken at the federal level to prevent racial discrimination[3] in employment decisions. In particular, this order attempted to provide federal agencies with sufficient power to enforce nondiscrimination policies (Gutman et al., 2011; Jones, 1977). As with earlier executive orders, 10925 applied to government employment and contracts. However, this executive order both increased accountability for contractors and created an agency with more enforcement power, the President's Committee on Equal Employment Opportunity. Specifically, government contractors were not only required to avoid discriminatory employment practices, but they also had to take affirmative action to ensure that applicants were treated without regard to race, creed, color, or national origin. For example, contractors needed to file regular compliance reports detailing their hiring and employment practices. Thus, this executive order was the starting point for the modern concept of affirmative action.

Finally, in an attempt to prevent the ineffectiveness of prior commissions, the President's Committee on Equal Employment Opportunity was allowed to initiate legal action by recommending suits against noncompliant or dishonest contractors to the U.S. Department of Justice. In addition, this committee could directly terminate government contracts with noncompliant contractors, forbid governmental agencies to sign future contracts with these contractors, and publish noncompliant contractors' names. Despite the number of enforcement options provided to the committee, not a single case was prosecuted under it (Gutman et al., 2011).

Civil Rights Act of 1964

The Civil Rights Act of 1964 was a landmark, multisection bill that prohibited discrimination across many aspects of Americans' lives. Although prior legislation largely focused on preventing racial discrimination in hiring, Title VII of the Civil Rights Act prohibits employment discrimination on the basis of race, color, religion, sex, or national origin. This section also established the EEOC. As with

[2]An executive order is a directive issued by the president of the United States that has the force of law. Such orders directly affect governmental agencies and their officials in addition to indirectly affecting private organizations that currently do business with, and want to continue doing business with, the federal government.

[3]Gender discrimination was introduced into legislation through the 1963 Equal Pay Act and the 1964 Civil Rights Act.

many of the preceding enforcement agencies, the EEOC initially lacked power to enforce Title VII; its options were limited to investigating complaints and seeking voluntary compliance. Because of restrictions on the EEOC and the long history of lax enforcement of fair employment regulations, most employers did not view Title VII as a threat (Guion, 1998; Jones, 1977).

President Johnson worked to strengthen Title VII enforcement among government contractors by issuing two executive orders. The first of these, Executive Order 11246 (1965), dissolved the Fair Employment Practice Committee and charged the U.S. Department of Labor with supervising the activities of government contractors. To this end, in 1965 the Office of the Secretary of Labor established the Office of Federal Contract Compliance Programs. Executive Order 11246, along with Executive Order 11375 (1967),[4] required federal contractors to include an equal opportunity claim in each contract, indicating the contractor's agreement not to discriminate on the basis of race, color, sex, creed, or national origin. Moreover, President Johnson perpetuated the prior administration's concept of affirmative action by delineating specific obligations for federal contractors to be in good standing with federal agencies (Gutman et al., 2011). In particular, contractors had to perform minority utilization in all job categories to determine whether there were fewer minorities or women in a particular job group than would reasonably be expected on the basis of their availability (Executive Order 11246, 1965). Also, contractors had to establish goals and timetables to correct any deficiencies. Finally, contractors had to develop data collection systems and reporting plans to document their progress in achieving these goals.

Equal Employment Opportunity Act of 1972 and Civil Rights Reform Act of 1978

Despite the optimism that accompanied it, Title VII did little to remedy discriminatory employment practices. Indeed, by 1972, the EEOC faced a backlog of more than 30,000 complaints, and the Justice Department was accused of dragging its feet in bringing Title VII suits. The situation was further complicated by the plethora of agencies involved in enforcing various aspects of EEO law. For example, the Department of Labor enforced the Equal Pay Act (1963) and the Age Discrimination in Employment Act (1967), and the U.S. Civil Service Commission and the U.S. Civil Rights Commission were responsible for ensuring EEO for federal employees (Guion, 1998). All of these agencies independently developed regulations and published guidelines (Guion, 1998), and at times the regulations from one agency contradicted the regulations from another.

The Equal Employment Opportunity Act of 1972 emerged out of this climate of confusion, disappointment, and disenchantment. This act amended Title VII by extending the coverage of the EEOC to smaller employers[5] as well as to state and local governments (Jones, 1977). Moreover, the act provided the EEOC with direct enforcement powers such as the ability to file injunctions against noncompliant employers when conciliation efforts failed. Finally, the act established the Equal Employment Opportunity Coordinating Council (EEOCC), which was charged with maximizing enforcement efforts and efficiency by eliminating conflict, competition, duplication, and inconsistencies among departments, agencies, and branches of the government concerned with EEO. This council consisted of the secretary of labor, the chairman of the EEOC, the attorney general, the chairman of the U.S. Civil Service Commission, and the chairman of the U.S. Civil Rights Commission (Guion, 1998). Unfortunately, the EEOCC was not able to accomplish its mission and was considered a failure.

Given the ineffectiveness of the EEOCC, President Carter initiated the reorganization of the federal government's EEO enforcement programs in 1978. The Civil Rights Reform Act of 1978 eliminated the EEOCC and established the EEOC as the

[4]Executive Order 11246 (1965) originally did not cover gender discrimination. Executive Order 11375 (1967) amended this by extending discrimination protection to women.

[5]The employment practices of organizations with 15 or more employees who worked 5 days a week for at least 20 weeks in the current or preceding calendar year were now covered by EEOC (Equal Employment Opportunity Act, 1972).

principal federal agency in fair employment enforcement. The EEOC was charged with enforcing Title VII, the Equal Pay Act, and the Age Discrimination in Employment Act as well as ensuring EEO for federal employees. In 1978, in conjunction with the Civil Service Commission, the Department of Labor, and the Department of Justice, the EEOC issued the *Uniform Guidelines on Employee Selection Procedures* (hereinafter, the *Uniform Guidelines*).

In summary, this brief review of the history of modern EEO law highlighted several issues. First, legislation was not effective in changing business practices until it was coupled with agencies with direct enforcement power. Second, early attempts to resolve discriminatory business practices at the state level were largely ineffective. Third, previous attempts to separate responsibility for EEO regulation resulted in conflict among federal agencies and confusion among organizations trying to meet EEO regulations. With the publication of the *Uniform Guidelines* in 1978, a single voice was finally communicating to organizations regarding EEO policy. However, to what extent are the *Uniform Guidelines* consistent with good professional practice in the field of psychology? This issue is explored in the next section.

PROFESSIONAL STANDARDS FOR EMPLOYMENT TESTING

The development of the 1978 *Uniform Guidelines* was driven by Supreme Court decisions regarding employment testing and validation (*Albermarle Paper Co. v. Moody,* 1975; *Griggs v. Duke Power,* 1971). The general goal of the guidelines was to help support EEO throughout all employment decisions (e.g., hiring, promotion, training) and to protect individuals belonging to the categories covered by Title VII of the 1964 Civil Rights Act. Specifically, the *Uniform Guidelines* address two general concerns: (a) Does the test create discrepancies between subgroups (i.e., adverse impact), and (b) does the test improve the efficiency or safety of the business (i.e., validity)? These topics are explored in greater detail later in this chapter.

The *Uniform Guidelines* cover issues related to the enforcement of Title VII and Executive Order 11246 and have not been updated since 1978. Thus, the guidelines do not address more recent legislation or judicial decisions, nor have they been updated to reflect changes in professional practice or the scientific literature on selection and testing. Consequently, there is debate over the usefulness of the *Uniform Guidelines*, which has largely focused on the different viewpoints advanced in this document and current professional practices (see McCauley, 2011). Indeed, although the *Uniform Guidelines* are one source providing information regarding validation standards, professionals also rely heavily on two more contemporary documents: the joint *Standards for Educational and Psychological Testing* (*Standards;* AERA et al., 1999), described in greater detail in Chapter 13 in this volume and the *Principles for the Validation and Use of Personnel Selection Procedures* (*Principles;* SIOP, 2003).

APA issued the first version of testing standards, titled *Technical Recommendations for Psychological Tests and Diagnostic Techniques,* in 1954. But later revisions were titled *Standards for Educational and Psychological Testing* (hereafter the *Standards*). It was designed to reflect scholarly consensus on test construction, evaluation, documentation (e.g., validity, reliability) and fairness in testing (e.g., language difficulties, disabilities) for all psychological and educational measurement. The *Technical Recommendations* were developed to provide guidelines for test takers and test users to encourage ethical use of tests and ethical testing practices. After the release of this first version of the *Standards,* APA joined with AERA and NCME to jointly revise the document. The *Standards* have been revised four times, with the most current revision published in 1999. A new revision was forthcoming at the time this chapter was being written and will soon be available.

In 1975, Division 14 of the American Psychological Association—then the Division of Industrial–Organizational Psychology, which would later become SIOP—published its own set of standards for employment testing, the *Principles*. The primary purpose of the *Principles* was to establish the perspective of industrial and organizational psychologists with regard to employment testing and assessment, especially validation (Jeanneret, 2005).

The *Principles* have been revised four times, with the most recent version published in 2003. Neither the *Principles* nor the *Standards* interpret legislation or judicial decisions related to testing or equal employment practices. However, both are used as valuable resources for understanding and developing valid selection tools. Despite the fact that both sets of guidelines have been updated to reflect current professional standards and practices, the older *Uniform Guidelines* still tend to be referenced more frequently in employment litigation (Jeanneret, 2005).

In the next section, important issues that need to be addressed when using tests in organizations are discussed. First considered is how test fairness is conceptualized in a legal sense. Next discussed are both legal and professional standards regarding applied testing. Finally, the shifting-burden-of-proof model, which is the decision process used when assessing an EEO court case, is reviewed. When discussing each of these topics, the perspectives of the *Uniform Guidelines,* the *Standards,* and the *Principles* are addressed.

IS THIS TEST CAUSING PROBLEMS? THE CONCEPT OF ADVERSE IMPACT

How does one know when the use of a test in organizations is problematic? One legal guideline for the fairness of organizational testing is the presence—or absence—of adverse impact, which is discussed in this section. Adverse impact occurs when the use of a test or some employment practice has differential consequences for two or more subgroups (e.g., racial, gender, religious). Discrimination under Title VII can be conceptualized under several different models. This discussion focuses on adverse impact because it is the primary model under which issues relevant to the use of tests and assessments in organizational settings emerge; interested readers are encouraged to consult Gutman (2005) and Gutman et al. (2011) for a comprehensive discussion of other litigation models.

As indicated, adverse impact (also known as *disparate impact*) is concerned with the consequences of the use of an employment practice (e.g., a test or a battery of tests). It is established by computing the impact of the employment practice on some minority subgroup and comparing it with the impact of the practice on the majority subgroup. For example, adverse impact can be said to exist if the percentage of African Americans passing an exam is sufficiently lower than the percentage of Caucasians passing the exam.

Two general techniques are used for determining whether two or more pass rates are sufficiently different to be labeled adverse impact (EEOC et al., 1978, Section 1607.3D). One technique is known as the four-fifths or 80% rule. According to this rule, adverse impact exists if the percentage of the minority subgroup passing the employment procedure is less than 80% of the ratio of the majority subgroup. The four-fifths rule was originally developed by the California Technical Advisory Committee on Testing in the early 1970s (Biddle, 2006) and was later codified in the *Uniform Guidelines* as a practical method for establishing adverse impact.

To apply this definition to the assessment of adverse impact, it is important to identify the majority subgroup. When the Civil Rights Act of 1964 was originally written, the majority subgroup was believed to be Caucasian and male, because the Civil Rights Act was interpreted as protecting certain groups from discrimination (e.g., women, African Americans). However, in *Regents of the University of California v. Bakke* (1978), the Supreme Court clarified that the belief that there are protected groups is incorrect. In particular, the Supreme Court determined that the Civil Rights Act prohibits all discrimination on the basis of race, color, religion, sex, or national origin. There are no protected groups because that term implies that there are groups that can be discriminated against and groups that are universally favored. Indeed, men or Caucasians can file discrimination lawsuits just as can women and African Americans.

Consequently, the majority subgroup in the adverse impact analysis is determined on the basis of composition in a job, organization, or industry. For example, if an occupation typically employs more women than men, then the majority subgroup for adverse impact analyses would be women. This context-specific approach to identifying the majority subgroup permits computation of adverse impact for reverse discrimination cases (e.g., men claiming

discrimination in a historically female-dominated job). Finally, the impact of an employment procedure does not have to be computed for every single subgroup in an applicant pool or organization. Only those subgroups that constitute at least 2% of the applicant pool are large enough to be used to determine adverse impact (EEOC et al., 1978).

The second method for determining whether the passing rates of two or more subgroups substantially differ is to conduct various statistical tests. The possibility of using statistical tests was mentioned in the *Uniform Guidelines* (EEOC et al., 1978, Section 1607.3D), which indicated that adverse impact can be inferred when there are statistically significant differences between minority and majority passing rates, provided the sample size is sufficient to permit meaningful analysis. A variety of statistical tests can be conducted, such as Fisher's exact test or the *Z* test for difference in proportions (also known as the *2 standard deviations test*). Specifically, the 2 standard deviations test is computed by using the following formula (Morris & Lobsenz, 2000):

$$Z_D = \frac{\left(PR_{\min} - PR_{\text{maj}}\right)}{\sqrt{PR_T(1 - PR_T)\left(\frac{1}{N_{\min}} + \frac{1}{N_{\text{maj}}}\right)}}, \quad (38.1)$$

where $PR_{\min}$ is the pass rate for the test for the minority group, PR_{maj} is the pass rate for the majority group, PR_T is the pass rate for the total sample, $N_{\min}$ represents the number of minority group applicants, and N_{maj} represents the number of majority applicants (Fleiss, 1981; Office of Federal Contract Compliance Programs, 1993). If the absolute value of the obtained Z_D is higher than 1.96, the pass rates are significantly different and adverse impact can be declared.

An important point regarding the four-fifths rule and the two aforementioned statistical tests for determining adverse impact necessitates further discussion. Regardless of how it is determined, adverse impact is a reflection of how a test is used and not an inherent property of a test. Specifically, some cut score must be determined so that the percentage of people passing a test can be determined for each subgroup. When there are subgroup distribution differences, the organization's choice of a cut score affects the likelihood of finding adverse impact. If the organization chooses a difficult cut score, the likelihood of finding adverse impact on such a test may increase. However, adverse impact may decrease or even disappear if an easier cut score is chosen for the very same test. Thus, adverse impact reflects how a test is used. Table 38.1 demonstrates potential inconsistencies between the four-fifths rule and statistical significance tests as a function of sample size.

Finally, it is important to note that demonstrating adverse impact is not proof that a test is discriminatory. Rather, demonstrating adverse impact simply establishes a *prima facie* case (*prima facie* is Latin for "on its first appearance"). Establishing a *prima facie*[6] case triggers further investigation into the employment practice in question (Guion, 1998; Gutman et al., 2011; Hanges, Aiken, & Salmon, 2011). It is at this stage that the court becomes interested in the psychometric quality of the employment practice and the validity of the inferences drawn from the practice. Indeed, organizations are not legally required to have any validity information regarding their employment procedures if these procedures do not exhibit any adverse impact (EEOC et al., 1978). However, once adverse impact has been found, then assessment quality and business necessity for the procedure has to be provided.

PSYCHOMETRIC QUALITY: RELIABILITY

Once adverse impact has been demonstrated, attention is turned to the quality of the employment procedure. Questions such as "Have adequate precautions been taken to minimize bias?" "Does the

[6]A second approach to establish a *prima facie* case is to demonstrate disparate treatment. Unlike adverse impact, which requires multiple observations, disparate treatment can be used with a single person or a small group of people. Disparate treatment is established by successfully arguing the following in court: (a) The person or persons belong to a minority subgroup in which minority subgroup status is consistent with the previous discussion; (b) the person or persons applied and was qualified for the job; (c) despite qualifications, the person was rejected; and (d) after rejection, the job stayed open, or the company looked for applicants with similar qualifications, or the company filled the job with someone having the same or lower qualifications as the person or persons filing the complaint. If the plaintiff is able to provide evidence to convince the judge that these four conclusions are appropriate, then a *prima facie* case is established.

TABLE 38.1

Comparison of Four-Fifths Rule With Statistical Test Definition of Adverse Impact (AI)

Applicants		Selected		Passing ratio			AI			
Majority	Minority	Majority	Minority	Majority	Minority	Total	Four-fifths rule	Decision	Z_D	Decision
25	10	20	5	.800	.500	.714	0.625	AI	−1.775	No AI
50	25	40	20	.800	.800	.800	1.000	No AI	0.000	No AI
100	100	99	98	.990	.980	.985	0.990	No AI	−0.582	No AI
200	200	118	95	.590	.475	.533	0.805	No AI	−2.305	AI

factor structure support the way the test is used?" and "Is there sufficient reliability to meaningfully make judgments with this test?" are asked. This domain is clearly one in which most psychologists would be very comfortable working.

Traditionally, actual employment decisions based on the tests were made from a rank-ordering perspective. That is, individuals who scored highest on a given test were believed to be better qualified than those who scored lower, regardless of the magnitude of difference between their scores. For example, suppose that for a particular employment test, one group of individuals scored 95 on the test and another group of individuals scored 96. The traditional rank-ordering model indicates that, on average, the latter group of individuals will outperform the former group of individuals. Indeed, if this strategy is applied in the long run, over a large number of applicants, selecting the applicants with the highest scores will yield the most economic utility (see Cascio, Outtz, Zedeck, & Goldstein, 1991). However, some researchers have questioned the extent to which organizations really operate in the long run and whether any particular employment test is sufficiently reliable to make such fine distinctions on the latent construct of interest (Cascio et al., 1991).

To address concerns with test reliability, Cascio et al. (1991) proposed the concept of test banding. According to classical test theory, observed score variance is a function of true score variance and random error variance (Nunnally, 1978; see also Chapter 2, this volume). As the degree of random error variance decreases, the test's reliability increases. Although theoretically possible, it is not possible in practice to ever have a perfectly reliable test. That is, random discrepancies between the observed score and the person's true potential will always occur. Test banding was proposed as a way to take this error into account when making selection decisions.

The variability in obtained test scores when people take two or more parallel tests, also known as the standard error of difference (*SED*), can be directly computed from the reliability of the test as shown in the following formula:

$$SED = S_x \sqrt{(1 - \rho_{xx'})}\ \sqrt{2}, \tag{38.2}$$

where S_x is the standard deviation of the test and $\rho_{xx'}$ is the reliability of the test. The width of the band is then obtained by multiplying the standard error of difference by some critical value (e.g., 1.96). Test bands are implemented in multiple stages. First, all the test scores are put in rank order. Then, one bandwidth is subtracted from the top-ranked test score. All applicants whose test scores fall within that bandwidth are considered to have latent true scores that are not different from each other.

The concept of test banding was first introduced in the applied testing literature as a way to reach a compromise between economic utility (e.g., hiring the candidates most likely to perform highly) and concerns for workplace diversity (Cascio et al., 1991). Specifically, if minority candidates routinely scored lower than Caucasian candidates on employment tests, then the use of a rank-ordering strategy to make decisions tends to result in adverse impact. However, if test scores are banded, some minority

candidates may be in the top band and thus be eligible for employment.

Indeed, although the use of test banding may appear to be attractive as an affirmative action procedure, legal decisions regarding the use of banding procedures have not always been supportive of this strategy. For example, banding has been upheld as an appropriate affirmative action procedure when decisions about who to hire within a band are not made solely on the basis of applicant race (e.g., *Chicago Firefighters v. City of Chicago*, 2001; *Bridgeport Guardians, Inc. v. City of Bridgeport*, 1991; *Jefferson County and Loeser v. Zaring and Hord*, 2002; *San Francisco Fire Fighters Local 798 v. San Francisco*, 2006). Use of banding as an affirmative action procedure has also been upheld when it is used as a temporary solution to remedy past racial injustices (e.g., *Officers for Justice v. the Civil Service Commission of the City and County of San Francisco,* 1992). In contrast, promotions awarded on the basis of a banding procedure were deemed improper in *Massachusetts Association of Minority Law Enforcement Officers v. Gerald T. Abban and Others* (2001). However, the rationale behind this ruling was due both to the focus on race in determining promotions and to a lack of evidence supporting the statistical generation of, and theoretical importance of, the banding procedure used.

It should be noted that test banding has generated considerable controversy in the scientific literature (Bobko & Roth, 2004). Although it is not a new concept, debate about banding seems to have been generated largely by the suggestion that psychometric theory is useful in establishing bandwidth. Proponents of psychometric-based banding strategies believe that banding accounts for error in measurement and can provide a more objective method of grouping individuals for employment decisions (Campion et al., 2001; Cascio et al., 1991). Detractors of these procedures voice concern over the inconsistency of such a strategy with the observed linear relationships between variables (for more on this discussion, see Aguinis, 2004; Campion et al., 2001). Moreover, although test banding is forwarded as a potential method to reduce adverse impact, a Monte Carlo simulation conducted in 1991 by Sackett and Roth demonstrated that test banding does not reduce adverse impact unless it is coupled with a within-band minority group preference selection strategy. Further complicating the issue, several different methods of banding have started to appear in the literature, and not all of these methods are rooted in the concept of test unreliability (e.g., Aguinis, 2004; Aguinis, Cortina, & Goldberg, 1998; Hanges & Gettman, 2004; Hanges, Grojean, & Smith, 2000).

In sum, the courts have generally upheld the use of banding in employment decisions, provided that race is not the sole criterion used to select applicants from within a band. Interestingly, some scholars have argued that these rulings may have limited the utility of one of the key motivations to use test banding—the desire to improve workforce diversity (Campion et al., 2001). However, the Supreme Court has not yet weighed in on the issue of banding (Barrett, Doverspike, & Arthur, 1995). Consequently, the use of banding in employment decisions may be further modified as case law builds in this area.

Although the *Uniform Guidelines* do not specifically address issues related to reliability, reliability is defined and addressed by both the *Standards* and the *Principles* (Jeanneret, 2005; see also Chapter 2, this volume). Both sets of professional standards point to the importance of assessing the consistency of scores across various sources of error, including time, raters, and items (SIOP, 2003, p. 70). The *Principles* hold that both the reliability of the test scores and the validity of inferences based on the test results should be determined (SIOP, 2003, p. 60).

APPROPRIATENESS OF INFERENCES: VALIDITY

In addition to examining a test's psychometric properties, important questions arise regarding whether the test is a necessary business practice. The Supreme Court explicated this standard in *Robinson v. Lorillard Corp.* (1971). Specifically, the ruling indicated that business necessity does not encompass an organization's ability to express some rationale or provide some justification for the challenged practice. Rather, organizations need to address whether "there exists an overriding legitimate business purpose such that the practice is necessary

for the *safe* and *efficient* [emphasis added] operation of the business."

In the scientific literature, *validity* refers to the appropriateness of inferences derived from a given test (SIOP, 2003), and it is not believed to be an inherent property of a test. In other words, a test can provide valid inferences about individuals in a clinical setting, but the same test may not provide valid inferences in another setting (e.g., selecting police officers). Demonstrating that a test provides appropriate inferences regarding current skill levels, subsequent job performance, or job potential speaks directly to the heart of the *Robinson v. Lorillard Corp.* business necessity requirement. Validity evidence demonstrates that the employment practice in question is connected to the efficient (and sometimes even the safe) operation of an organization.

Validity as Codified in the *Uniform Guidelines*

According to the *Uniform Guidelines,* there are three types of validity, and different types of evidence are needed to support each of type. Thus, the *Uniform Guidelines* use what has been called the *trinitarian* perspective on validity (Guion, 1980; Landy, 1986). Specifically, the *Uniform Guidelines* indicate that validity can be supported by (a) establishing a relationship between the scores on the procedure in question and job performance (i.e., criterion validity); (b) a professional assessment of the overlap in content between the employment procedure and the job itself (i.e., content validity); or (c) a demonstration that the procedure measures a construct that is important for job performance (i.e., construct validity). The *Uniform Guidelines* provide considerable detail explicating what needs to be met for each kind of validity (e.g., a job analysis must be done for content validity; the sample in a criterion-related validity study should be similar to the relevant labor market).

Not only did the *Uniform Guidelines* specify three different types of validity, they also specified when each type of validity should be collected. Thus, the *Uniform Guidelines* stated that inferences about mental processes

> cannot be supported solely or primarily on the basis of content validity. Thus, a content strategy is not appropriate for demonstrating the validity of selection procedures which purport to measure traits or constructs, such as intelligence, aptitude, personality, commonsense, judgment, leadership, and spatial ability. (EEOC et al., 1978, § 1607.14.C(1))

However, content validity studies are recommended when the organization is developing work samples or measures of competencies necessary for successful performance.

Because the *Uniform Guidelines* discuss three types of validity, the question has arisen—in court—as to whether the *Uniform Guidelines* "prefer" one form of validity over another. Indeed, they do not. Instead, as previously discussed, the authors of the guidelines felt that each form of validity evidence would be uniquely useful in a particular situation. This viewpoint has been upheld in court, as can be seen in *Gillespie v. State of Wisconsin* (1986).

With respect to criterion-related validity, one important issue to note is that the *Uniform Guidelines* were written when researchers believed in the situational specificity hypothesis, which was largely false. That is, researchers thought that a test's job relatedness did not necessarily transfer from one situation to another even though the exact same job was being performed in both locations. As a result, the *Uniform Guidelines* have been interpreted as emphasizing the need for separate validity studies in each location to prove the validity of the same employment practice for the same job in different locations. However, although the *Uniform Guidelines* were developed under the assumption of situational specificity, they were also written to be interpreted in light of current scientific evidence. Although some scholars have argued that the *Uniform Guidelines* exclude the use of validity generalization arguments (McDaniel, Kepes, & Banks, 2011; see also Chapter 4, this volume), validity generalization arguments have been successfully upheld twice thus far in court. First, in *Williams et al. v. Ford* (1999), Ford was able to support the validity of its selection test by using a combination of a criterion-related validity study conducted with its workforce and a

meta-analysis of similar tests used by other employers in similar jobs. Second, in *Association of Mexican-American Educators et al. v. the State of California* (2000), the State of California successfully argued for the validity generalization of findings from two prior studies of the test in question in similar jobs. Specifically, the court found that, despite differences in the abstraction of the measurement of skills, the prior studies provided sufficient evidence for validity with respect to the current use of the test. Although one might argue that the courts in these cases simply did not accept the guidelines as determinative, the rulings in both explicitly stated that the guidelines should be given deference and assessed the validation studies in question against the requirements for validation outlined in the guidelines.

Indeed, careful reading of the *Uniform Guidelines* reveals that although it does not explicitly include the words *validity generalization,* it specifies a validity transportability procedure that could be used to import validity information from one setting to another (Biddle & Nooren, 2006). Unfortunately, the validity transportability procedure discussed in the *Uniform Guidelines* is limited only to transporting the exact same test for the exact same job across different locations. Validity generalization, however, is a more general procedure in that it explores the robustness of inferences across different locations, different jobs, and even different tests (given evidence that the tests measure the same psychological construct).

Finally, the *Uniform Guidelines* discuss the importance of conducting studies of fairness, when technically feasible. Readers interested in a more thorough discussion of test fairness may want to consult Volume 3, Chapter 27, this handbook. Analyses of fairness generally consist of performing the Cleary (1968) moderated regression fairness model to determine whether the test has differential prediction of a given outcome as a function of race, color, religion, sex, or national origin. Specifically, differential prediction occurs when the same test score has different meanings (i.e., different outcomes) for individuals from different subgroups. There are three types of differential prediction according to this model: intercept bias, slope bias, and the combined intercept–slope bias. Regardless of the specific type, evidence of differential prediction is problematic because the same test score (e.g., 90) has a different meaning for one subgroup (e.g., adequate job performance) than for the other (e.g., exceptional job performance). As with many of the issues reviewed in this chapter, differential prediction has been thoroughly discussed in the industrial and organizational psychology literature. Some researchers have claimed that differential prediction does not happen or, when it does, favors minority subgroup members (Hunter, Schmidt, & Hunter, 1979). Others have asserted that differential prediction occurs, but not universally (Van Iddekinge & Ployhart, 2008). In other words, differential prediction might emerge in certain subgroup comparisons for certain criteria, but not for all subgroup comparisons across all criteria.

It should be noted that the Cleary (1968) approach to assessing differential validity assumes that the criterion variable is not sensitive to discriminatory factors (e.g., rater prejudice, opportunity bias). To the extent that such factors affect subgroup differences on the criterion, the accuracy of the differential prediction statistical analysis will be affected (Saad & Sackett, 2002). Thus, it is imperative that researchers choose their criterion variable carefully when conducting these analyses (Van Iddekinge & Ployhart, 2008).

The *Principles* and the *Standards* also devote a great deal of time to discussions of validity. Although the *Standards* and the *Principles* are consistent in their treatment of validity (Jeanneret, 2005), compared with the *Standards*, the *Principles* provide more information on validation, job analysis, and data analysis. For example, the *Principles* discuss validity generalization, including synthetic–job component validity[7] evidence, meta-analysis, and cut scores. In contrast to the *Uniform Guidelines,* both the *Standards* and the *Principles* view validity as

[7]*Synthetic validity,* also known as *job component validity,* is used when a single organization does not have a sufficient sample size in a single job to conduct a criterion-related validity study. Synthetic validity is conducted by analyzing multiple jobs on their component skills and identifying a family of jobs that require the test's skill set. Employees in these jobs take the test, and their scores are correlated against some criterion. The validity of an entire test battery is synthesized by combining the separate validities obtained across multiple validity studies.

a unitary construct, discuss convergent and discriminant validity, and consider fairness and bias from multiple perspectives. Interested readers are encouraged to refer to Jeanneret (2005) for a more detailed comparison of the professional standards on the topic of validity.

Legal and Judicial History Regarding Validation Evidence

In addition to the validation guidelines provided in the *Guidelines*, *Standards*, and *Principles*, legal decisions have also shaped the standards by which validation evidence is judged. We review several salient cases on validation guidelines in the sections that follow.

***Albemarle Paper Co. v. Moody* (1975).** The *Albemarle Paper Co. v. Moody* case was important in establishing guidelines for validity evidence, because the ruling explicitly discussed the quality of the validity study provided by the Albemarle Paper Company. On the eve of the trial, the company hired an industrial psychologist who spent approximately half a day completing a concurrent criterion-related validity study. The psychologist did not perform a job analysis and did not have a sufficient sample size to conduct the criterion-related study in any one job. So he created a sufficiently large sample to conduct the study by grouping jobs together solely on the basis of the proximity of their line of progression.[8] The dependent variable in the validity study was obtained by asking each supervisor to independently rank their subordinates. Unfortunately, no information was provided regarding the criteria to consider when making these rankings.

The Supreme Court, relying on guidelines published by the EEOC and the 1974 *Standards* (APA, AERA, & NCME, 1974), found that the Albemarle validation design was deficient. Specifically, the court found that there was no way to precisely discern what standards the individual supervisors used to rank their employees, or whether the supervisors were even using the same standard. In addition, the sample used in the validation study mainly came from jobs near the top of progression line. Inconsistent with the 1970 EEOC guidelines, the study's sample was not representative of the relevant labor market for entry-level jobs. Finally, the psychologist did not conduct test fairness studies (i.e., he did not test whether the tests were differentially valid for the subgroups), nor did he argue that such an analysis was technically infeasible. In terms of the implications for testing and assessment, the *Albemarle Paper Co. v. Moody* case highlights that the court will examine the technical details of a validity study, and if the validity study is deemed inadequate, the defendant is in trouble.

***Washington v. Davis* (1976).** The Washington, DC, police department used a verbal ability test, called Test 21, for selection and promotion purposes. Test 21 excluded 4 times as many African American candidates as Caucasian candidates, with 57% of African American candidates and 13% of Caucasian candidates failing the test between 1968 and 1971. Two African American police officers alleged that the department's recruiting policy was racially discriminatory. Instead of filing the complaint under Title VII, the officers wanted the test declared unlawfully discriminatory and thus a violation of the due process clause of the Fifth Amendment to the U.S. Constitution. The department countered with validity evidence showing that Test 21 predicted performance in the department's officer training programs. Furthermore, the department argued that it had a program to recruit African American applicants and that the African American population between ages 20 and 29 was proportionate to the number of African American police on the force.

This case is important for a number of reasons, some of which are discussed later in this chapter. However, this case brought to light issues bearing on the legal guidelines for validity. Specifically, the defendant provided validity evidence showing that Test 21 predicted performance in the department's officer training programs. Previously, only actual job performance was used as a criterion to demonstrate business necessity. However, because the court ruled in favor of the defendant in this case (see explanation later in this chapter), performance in training became an acceptable criterion for

[8]Although this strategy might have been supported if the psychologist had collected job analysis information and clustered jobs on the basis of overlap in the knowledge, skills, and abilities required to do them, this information was not collected.

establishing business necessity. More technically, given safety issues involved in the job, the court argued that it was sufficient to compare test scores with the content-valid prerequisite training and not necessary to compare the test with actual job performance as a police officer (Gutman, 2005, p. 28).

In summary, in this section the authors discussed the importance of providing validity information to demonstrate that an employment practice in question meets the business necessity standard specified in *Robinson v. Lorillard Corp.* Moreover, different conceptualizations of validity used in the guidelines and professional standards were addressed, and several recommendations for validity studies made by the *Uniform Guidelines* were argued to be inconsistent with current professional standards. It should be noted that even though these discrepancies exist, the *Uniform Guidelines* still influence governmental agencies and other organizations. Thus, it is imperative that psychologists understand not only the currently accepted professional standards but also the boundaries placed on organizations by these standards. In the next section, the standards used to decide EEO court cases are reviewed. This model, which developed through the evolution of case law, is called the *shifting-burden-of-proof model.*

SHIFTING BURDENS AMONG PLAINTIFFS AND DEFENDANTS

The process of deciding an adverse impact case follows what is called the *shifting-burden-of-proof model.* This is a decision process that the judge uses when weighing the evidence presented in a case. In this decision model, the burden of proof first rests with the plaintiff. The plaintiff must present convincing evidence demonstrating that adverse impact has occurred. If the judge is convinced by the plaintiff's evidence, the burden of proof shifts to the defendant. The defendant's burden is to demonstrate that the employment practice is sound and has validity. If the judge is convinced that the practice has validity, the burden of proof shifts back to the plaintiff. At this time, the plaintiff must furnish evidence that the employment practice is a pretext for discrimination. Specifically, the plaintiff may do this by demonstrating that there are other employment practices that are equally valid but result in less adverse impact. In other words, the plaintiff has to argue that the defendant could easily have accomplished the business purpose with a procedure that would have produced less harm to the particular subgroup in question. It should be noted that this orderly procession of burden shifting only takes place in the judge's mind. If one observed an actual trial, one would see evidence and counterevidence touching on all of these issues presented throughout the trial.

Legal and Judicial History of the Shifting-Burden-of-Proof Model

Unlike standards of validation, the shifting-burden-of-proof model arose exclusively from case law. We review three landmark cases in the development of this model in the sections that follow.

***Griggs v. Duke Power* (1971): Intent versus consequences.** In the first case prosecuted under Title VII, the Duke Power Company was accused of practicing discriminatory hiring and assignment practices. Before the passage of the Civil Rights Act of 1964, the company openly discriminated on the basis of race by enforcing a policy that African Americans could only work in the Labor plant. Among the plants run by Duke Power, this plant had the lowest-paying jobs. In 1955, Duke Power instituted a requirement that employees transferring between plants had to have a high school diploma—something accomplished primarily by Caucasians rather than African Americans in North Carolina at that time. When Title VII became effective, the company added two professionally developed tests, the Wonderlic Personnel Test and the Bennett Mechanical Comprehension Test, to their requirements for employees to transfer plants. These testing requirements effectively prohibited the transfer of African American employees into more lucrative jobs in the company. Early judgments favored the company, with the District Court and Court of Appeals ruling that the company had not violated Title VII because the testing requirements were applied equally to Caucasian and African American employees. In addition, the lower courts found that the plaintiff failed to establish that the company acted out of discriminatory intent.

The Supreme Court overruled the lower courts' judgments, asserting that the absence of discriminatory intent does not justify selection procedures that are not job related. This ruling was important in setting the standards for determining adverse impact because it highlighted that the consequences of employment practices are the principal concern; it is not necessary to prove that the organization intended to discriminate when pursuing a case under the adverse impact model. Moreover, the court defined the shifting-burden-of-proof model in this case by affirming that the touchstone of the Civil Rights Act is business necessity; after the plaintiff establishes a *prima facie* case, the defendant must show that the challenged practice is related to the job in question. *Griggs v. Duke Power Co.* is significant for workplace testing practices in that it highlights the importance of monitoring the consequences of testing rather than relying on good intentions as well as the importance of ensuring that employment tests are job related.

***Washington v. Davis* (1976).** As discussed previously, the Washington, DC, police department used a verbal ability test, called Test 21, for selection and promotion purposes. This test excluded 4 times as many African American candidates as Caucasian candidates. The critical point to focus on in this case is that the complaint was not filed under Title VII; rather, the test was claimed to be unlawfully discriminatory because it violated the due process clause of the Fifth Amendment to the U.S. Constitution.

The Supreme Court ruled that a case filed under constitutional rule carries a heavier burden for the plaintiff than do cases filed under Title VII. Specifically, cases filed under constitutional rule require evidence of purposeful discrimination, whereas Title VII cases do not. This ruling helped to solidify standards for establishing a *prima facie* case under Title VII relative to other models of discrimination.

***Connecticut v. Teal* (1982): Bottom line.** The *Uniform Guidelines* originally advocated a bottom-line approach to assessing adverse impact. In other words, if an organization required applicants to take a battery of tests to make an employment decision, the bottom-line approach would say that if there is no adverse impact on the final employment decision, a *prima facie* case would not be established. The adequacy of the bottom-line defense was tested in *Connecticut v. Teal*.

Four African American employees of the Department of Income Maintenance of the State of Connecticut were provisionally promoted to the position of welfare eligibility supervisor. To attain permanent status, the employees had to complete a multistep selection process. The first step of the process was a written test, which had a pass rate of 54.17% for African Americans and 79.54% for Caucasians. The respondents in this case alleged that the written test violated Title VII because it excluded African American applicants in disproportionate numbers and was not job related. More than a year after filing the complaint, and approximately 1 month before the trial, the four employees were promoted on the basis of an eligibility list generated by the written test. The department instituted affirmative action plans, promoting 22.9% of African American employees and 13.5% of Caucasian employees. The department justified the affirmative action program by arguing that it maintained the bottom-line statistics for the promotion system.

The Supreme Court ruled that bottom-line defense was unacceptable and the individual components of a selection battery could be investigated even when the overall employment decision does not have adverse impact. In other words, if adverse impact is found in any component of a selection system, the defendant must justify that component. In this decision, the Supreme Court differentiated fairness to a group from fairness to individuals. In particular, the justices said that it is unacceptable for the final decisions of a selection system to show no discrimination (i.e., fairness to subgroups) when actual people are being prevented from pursuing job opportunities by a discriminatory test subcomponent (i.e., fairness to individuals). Thus, this case warned organizations that tests can be targeted for investigation even if the final employment decisions do not exhibit adverse impact.

Challenging the Shifting-Burden-of-Proof Model

The shifting-burden-of-proof model established in *Griggs v. Duke Power* (1971) appeared to be well

established by the late 1980s. However, two cases then challenged the very structure of the model. In *Watson v. Fort Worth Bank and Trust* (1988), an African American female bank employee, Clara Watson, was denied promotion in favor of Caucasian employees on the basis of a multiple-component subjective assessment system, including ratings of job performance, interview performance, and past experience. The evaluation system as a whole showed adverse impact, but it was impossible to disaggregate the components of the system for further analysis.

Consistent with previous decisions, the Supreme Court ruled that subjective assessment systems, such as the one used by Fort Worth Bank, could be subjected to adverse impact analysis. However, in a move inconsistent with the shifting-burden-of-proof model, the Court ruled that the plaintiff was responsible not only for demonstrating adverse impact, but also with identifying the cause or causes of adverse impact within the selection system or proving that the selection components could not be disaggregated. Moreover, the Court ruled that when the burden shifts to the defendant, the employer only has to state a legitimate business purpose for the contested practice instead of presenting validity evidence. Thus, this decision reduced the defendant's burden by moving away from the *Robinson v. Lorillard Corp.* (1971) business necessity requirement while simultaneously increasing the plaintiff's initial burden.

The issues of subjective employment assessments, identification of the cause of adverse impact in multicomponent selection systems, and the change in the shifting-burden-of-proof model from *Watson v. Fort Worth Bank and Trust* (1988) reappeared in *Wards Cove Packing Co. v. Atonio* (1989). In the Alaskan Wards Cove Packing Company, a salmon cannery, jobs were classified as either skilled or unskilled. The skilled jobs were occupied primarily by Caucasian employees, whereas employees in the unskilled jobs were predominately Eskimo and Filipino (Gutman et al., 2011). The plaintiffs tried to establish a *prima facie* case by arguing that use of the overall selection system created adverse impact.

Consistent with *Watson v. Fort Worth Bank and Trust* (1988), the Supreme Court determined that by not identifying the specific selection procedure responsible for the disparities, the plaintiffs had not established a *prima facie* case. Moreover, the court found that the defendant's burden of proof requires them only to provide evidence that the targeted practice serves legitimate employment goals and not that the practice is job related.

Reinstating the Shifting-Burden-of-Proof Model: Civil Rights Act of 1991

Following the *Wards Cove Packing Co. v. Atonio* (1989) decision, Congress proposed a number of bills aimed at reversing or qualifying its effects, especially the changes that these decisions made to the *Griggs v. Duke Power Co.* (1971) shifting-burden-of-proof model. The 1990 Civil Right Act specifically targeted *Wards Cove,* addressing the issues of bottom-line effects on establishing *prima facie* cases, the shifting-burden-of-proof model, and the standards for business necessity. This bill was vetoed by President George H. W. Bush, and two additional congressional acts also failed before the passage of the 1991 Civil Rights Act. The 1991 act was intended to clarify issues related to causation, the shifting burden of proof, and business necessity. It also addressed the practice of adjusting selection test scores or cutoffs on the basis of class membership.

The 1991 act attempted, with varying degrees of success, to resolve the issues brought to light in the *Wards Cove Packing Co. v. Atonio* (1989) decision. First, regarding identification and causality, the act specifies that plaintiffs must specify which selection procedures caused adverse impact, except in cases in which the selection system cannot be separated into components for analysis. If the system cannot be separated, the entire employment system can be analyzed as a single practice. The act also reversed the *Wards Cove* decision and reestablished the *Griggs v. Duke Power Co.* (1971) shifting-burden-of-proof model. That is, the act specified that after the plaintiff's establishment of a *prima facie* case, the defendant must provide evidence that the targeted practice is related to the efficiency or safety of the business. Last, the act also addressed the topic of race norming, an unpopular practice involving the adjustment of scores or score cutoffs based on race or other protected EEOC categorization; the act expressly forbade this practice in employment tests.

In summary, in this section the authors discussed how the *Griggs v. Duke Power Co.* (1971) shifting-burden-of-proof model has been used to decide an EEO court case. Although the nature of the model faced challenges in the late 1980s, the Civil Rights Act of 1991 reestablished the original model. However, this act also opened up other possibilities such as the use of jury trials to decide these cases. So, as of the writing of this chapter, the full impact of the Civil Rights Act of 1991 has not completely played out. A number of legal and psychometric concerns that psychologists must face when developing and implementing tests in organizations have been presented. Two of these concerns—validity and reliability—are very familiar to psychologists. However, with respect to employment testing, the investigation of these concerns is inherently entwined with issues regarding the consequences of test use. For this reason, this chapter ends with a discussion of strategies used to reduce adverse impact while maintaining the validity of an employment practice.

REDUCING ADVERSE IMPACT

Strategies for minimizing adverse impact have long captured the attention of researchers, practitioners, and lawmakers. In their 2001 review, Sackett, Schmidt, Ellingson, and Kabin reported that several often-repeated suggestions for reducing adverse impact do not actually work. Among these are coaching programs, providing generous time limits, and attempting to improve applicant test-taking motivation. However, other methods are being attempted. Each of these methods is presented briefly, as is their legal standing; for a more in-depth review, see Gutman et al. (2001), Sackett et al. (2001), and Sackett and Wilk (1994).

One method of reducing adverse impact that has been attempted involves manipulating the content of a test by eliminating problematic items (i.e., items exhibiting differential item functioning or other undesirable patterns). However, the timing of when this analysis and item elimination is done (e.g., during test development vs. after test administration) is critical for successfully defending this tactic (Sackett & Wilk, 1994). Clearly, items may be freely eliminated for a variety of reasons during the test development phase (Sackett & Wilk, 1994). However, the legality of eliminating items becomes muddied when items or test components are eliminated after the test is administered to actual job applicants. In *Hayden v. Nassau County* (1999), Nassau County and the Department of Justice worked jointly to create a test with no adverse impact that complied with Title VII and the *Uniform Guidelines*. After two failed attempts, a new design committee developed a 25-component test that was administered to 25,000 applicants. Unfortunately, this test also produced adverse impact. The county used the applicant data to eliminate 16 of the 25 original test components, thereby creating a new nine-component test that had less adverse impact than the original longer test. However, a group of unsuccessful candidates challenged this shortened test by arguing that they would have been selected on the basis of the original test. The court rejected the challenge because, even though race was considered when determining whether to discard a particular test component, test component elimination was done in a race-neutral fashion. Nevertheless, the unique situation of this case, specifically the cooperation between Nassau County and the Department of Justice, makes it difficult to assess the implications of *Hayden* for other employers who eliminate test items (Gutman et al., 2011). In general, however, Sackett et al.'s (2001) review concluded that eliminating items showing differential item functioning was not effective in reducing adverse impact.

Perhaps one of the most extreme ways to reduce adverse impact is to discard the test altogether. In the recent *Ricci v. DeStefano et al.* (2009) case, the New Haven Civil Service Board discarded promotion exam results for firefighters after finding that the test produced adverse impact. The board did so on the good-faith belief that they would lose an adverse impact claim if sued by minority candidates. A group of 17 Caucasian firefighters and one Hispanic firefighter challenged the decision to discard the results and sued the board. The Supreme Court overruled lower court decisions and found that the board had acted inappropriately by discarding the test results because the board did not have strong evidence that the test was technically deficient. The board only discarded the results to avoid violating

the disparate impact clause of Title VII. In other words, the court found that the board discarded the test on the basis of race, despite the fact that it did not have evidence that anything was truly wrong with the test.

Another method of reducing adverse impact is to seek alternative testing procedures that minimize group differences, when such alternatives are available (Guion, 1998; Gutman et al., 2011). Indeed, the scientific literature supports the belief that noncognitive tests are useful and may reduce adverse impact (Goldstein, Yusko, Braverman, Smith, & Chung, 1998) and that expansion of the testing construct domain beyond cognitive ability (e.g., personality) reduces adverse impact (Sackett et al., 2001). The use of alternative testing procedures has been repeatedly called for in case law, legislation, and the *Uniform Guidelines* (Gutman et al., 2011). For example, the *Uniform Guidelines* state that given two procedures that serve the employer's legitimate interests and that are equally valid, the procedure with less adverse impact should be used. The requirement to use alternative tests and assessments that produce less or no adverse impact was also codified in the Civil Rights Act of 1991.

Finally, another method for reducing adverse impact is differentially weighting the information obtained from various predictors when making a selection decision. For example, the adverse impact of a selection procedure consisting of a cognitive ability test and a personality test will differ if one test is weighted more than the other. Specifically, less adverse impact will be found if the personality test is weighted more heavily than the cognitive ability test, whereas more adverse impact will be found if the opposite weighting scheme is used (Sackett & Ellingson, 1997). Of course, differential predictor weights affect the kinds of people that end up in an organization.

Various predictor weighting strategies have been attempted. For example, regression-based weights are developed on the basis of some empirical validity study and these weights maximize the prediction of the regression equation's dependent variable, which is usually job performance (De Corte, Lievens, & Sackett, 2008). Another approach is to establish minimum passing scores for all predictors and a policy of randomly selecting employees from among the pool of applicants who have successfully passed all the predictor minimum scores. Although the minimum-passing-score approach produces less adverse impact than the regression-based approach, the average quality of the employees hired under this strategy is also lower. A relatively new approach, called the *pareto-optimal approach,* was designed to find the optimal combination of predictor weights that simultaneously maximizes the prediction of performance while minimizing adverse impact (De Corte, Lievens, & Sackett, 2007). Simulations testing the utility of this technique have shown that it has substantial promise.

FINAL THOUGHTS

As discussed throughout this chapter, psychologists interested in using tests to assist organizational decision making have to meet not only their own professional standards but also legal standards with regard to test use. Similar to the *Principles* and the *Standards,* the *Uniform Guidelines* are also concerned with the soundness and validity of the test. However, the legal world is also concerned with the consequences of test use, something that is not a primary focus in the test construction literature (for an exception, see Messick's 1995 discussion of the consequential aspect of validity). It is clear that psychologists doing applied work cannot be effective without being aware of the legal standards, court cases, and legislation.

The authors of this chapter also believe that the psychologist as scholar personally benefits by understanding testing standards from both the legal and the psychological fields. The development and use of tests in employment scenarios are hotly debated within the scientific, practitioner, and legal communities. Indeed, major controversies remain over the legal definitions of *validity* (as codified in the *Uniform Guidelines*) and *test deficiency* as well as the use of banding in employment testing. There is also considerable debate within the field of psychology about how science should shape the law regarding, practice of, and future guidelines for employment testing. Moreover, the understanding of what influences tests and creates adverse impact is expanding

(Chapman, Uggerslev, Carroll, Piasentin, & Jones, 2005; Hausknecht, Day, & Thomas, 2004; Jenneret, 2005; Outtz, 2009). These new insights are the result of researchers working in applied settings with multiple constituents, who raise new questions that are then brought back into the science for answers. The individual scholar and the field of industrial and organizational psychology benefit from this interaction.

References

Age Discrimination in Employment Act of 1967, 29 U.S.C. 621 *et seq.*

Aguinis, H. (2004). *Test score banding in human resource selection: Legal, technical, and societal issues.* Westport, CT: Praeger.

Aguinis, H., Cortina, J. M., & Goldberg, E. (1998). A new procedure for computing equivalence bands in personnel selection. *Human Performance, 11*, 351–365. doi:10.1207/s15327043hup1104_4

Albermarle Paper Co. v. Moody, 422 U.S. 405 (1975).

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist*, *58,* 377–402.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests.* Washington, DC: Author.

American Psychological Association, Division of Industrial–Organizational Psychology. (1975). *Principles for the validation and use of personnel selection procedures.* Dayton, OH: Author.

Association of Mexican-American Educators et al. v. the State of California, 231 U.S. 572 (2000).

Barrett, G. V., Doverspike, D., & Arthur, W., Jr. (1995). The current status of the judicial review of banding: A clarification. *Industrial-Organizational Psychologist, 33,* 39–41.

Biddle, D. A. (2006). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing* (2nd ed.). Burlington, VT: Ashgate.

Biddle, D. A., & Nooren, P. M. (2006). Validity generalization vs. Title VII: Can employers successfully defend tests without conducting local validation studies? *Labor Law Journal, 57*, 216–237.

Bobko, P., & Roth, P. L. (2004). Personnel selection with top-score referenced banding: On the inappropriateness of current procedures. *International Journal of Selection and Assessment, 12*, 291–298. doi:10.1111/j.0965-075X.2004.00284.x

Bridgeport Guardians, Inc. v. City of Bridgeport, 933 F. 2d 1140 (1991).

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology, 54,* 149–185. doi:10.1111/j.1744-6570.2001.tb00090.x

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). The implications of six methods of score use in personnel selection. *Human Performance, 4,* 233–264. doi:10.1207/s15327043hup0404_1

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology, 90*, 928–944. doi:10.1037/0021-9010.90.5.928

Chicago Firefighters v. City of Chicago, 249 F. 3d 649 (2001).

Civil Rights Act of 1866, ch. 31, 15 Stat. 27. (1866).

Civil Rights Act of 1964, Pub. L. 88–352, 42 U.S.C. Stat. 253 (1964).

Civil Rights Act of 1991, Pub. L. 102–166, 105 Stat. 1071. (1991).

Civil Rights Reform Act of 1978, Pub. L. 95–454, 92 Stat. 1111 (1978).

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x

Connecticut v. Teal, 457 U.S. 440 (1982).

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380–1393. doi:10.1037/0021-9010.92.5.1380

De Corte, W., Lievens, F., & Sackett, P. R. (2008). Validity and adverse impact potential of predictor composite formation. *International Journal of Selection and Assessment, 16*, 183–194. doi:10.1111/j.1468-2389.2008.00423.x

Equal Employment Opportunity Act of 1972, Pub. Law 92–261, 86 Stat. 103. Retrieved from http://www.eeoc.gov/eeoc/history/35th/thelaw/eeo_1972.html

Equal Employment Opportunity Commission. (1970). Guidelines on employee selection procedures. *Federal Register, 35,* 12333–12336.

Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*, 38290–39315.

Equal Pay Act of 1963, Pub. L. 88–38, 77 Stat. 56. (1963)

Exec. Order No. 8802, 3 C.F.R. 957 (1941).

Exec. Order No. 9346, 3 C.F.R. 1280 (1943).

Exec. Order No. 10925, 3 C.F.R. 448 (1961).

Exec. Order No. 11246, 3 C.F.R. 339 (1965).

Exec. Order No. 11375, 3 C.F.R. 684 (1967).

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: Wiley.

Gillespie v. the State of Wisconsin, 771 U.S. 1035 (1986).

Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51*, 357–374. doi:10.1111/j.1744-6570.1998.tb00729.x

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*, 385–398.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Gutman, A. (2005). Adverse impact: Judicial, regulatory, and statutory authority. In F. J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 20–46). San Francisco, CA: Jossey-Bass.

Gutman, A., Koppes, L., & Vadonovich, S. (2011). *EEO law and personal practices* (3rd ed.). New York, NY: Routledge.

Hanges, P. J., Aiken, J. A., & Salmon, E. D. (2011). The devil is in the details (and the context): A call for care in discussing the *Uniform Guidelines*. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 562–565.

Hanges, P. J., & Gettman, H. (2004). A comparison of test-focused and criterion-focused banding methods: Back to the future? In H. Aguinis (Ed.), *Test score banding in human resource selection: Legal, technical, and societal issues* (pp. 29–48). Westport, CT: Praeger.

Hanges, P. J., Grojean, M. W., & Smith, D. B. (2000). Bounding the concept of test banding: Reaffirming the traditional approach. *Human Performance, 13*, 181–198. doi:10.1207/s15327043hup1302_4

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639–683. doi:10.1111/j.1744-6570.2004.00003.x

Hayden v. Nassau County 180 F. 3d 928 (1999).

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*, 721–735. doi:10.1037/0033-2909.86.4.721

Jeanneret, A. (2005). Professional and technical authorities and guidelines. In F. J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 47–100). San Francisco, CA: Jossey-Bass.

Jefferson County & Loeser v. Zaring & Hord, 91 U.S. 583 (2002).

Jones, J. E., Jr. (1977). The development of modern equal employment opportunity and affirmative action law: A brief chronological overview. *Howard Law Journal, 20*, 74–99.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192. doi:10.1037/0003-066X.41.11.1183

Massachusetts Association of Minority Law Enforcement Officers v. Gerald T. Abban & others, 434 Mass. 256 (2001).

McCauley, C. (Ed.). (2011). *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*(4).

McDaniel, M. A., Kepes, S., Banks, G. C. (2011). The *Uniform Guidelines* are a detriment to the field of personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 494–514.

McDonald v. Santa Fe Trail Transportation Co., 427 U.S. 273 (1976).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*, 89–111. doi:10.1111/j.1744-6570.2000.tb00195.x

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Office of Federal Contract Compliance Programs. (1993). *Federal contract compliance manual* (SUDOC No. L 36.8: C 76/993). Washington, DC: U.S. Department of Labor, Employment Standards Administration, Office of Federal Contract Compliance Programs.

Officers for Justice v. the Civil Service Commission of the City and County of San Francisco, 979 F. 2d 721 (1992).

Outtz, J. L. (2009). *Adverse impact: Implications for organizational staffing and high stakes selection.* New York, NY: Routledge.

Regents of the University of California v. Bakke, 438 U.S. 265 (1978).

Ricci v. DeStefano et al., 530 F. 3d 87 (2009).

Robinson v. Lorillard Corp., 444 F. 2d 791 (1971).

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment- oriented personality measures. *Journal of Applied Psychology, 87*, 667–674. doi:10.1037/0021-9010.87.4.667

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707–721. doi:10.1111/j.1744-6570.1997.tb00711.x

Sackett, P. R., & Roth, R. (1991). A Monte Carlo examination of banding and rank order methods of test score use in personnel selection. *Human Performance, 4*, 279–295. doi:10.1207/s15327043hup0404_3

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education—Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302–318. doi:10.1037/0003-066X.56.4.302

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*, 929–954. doi:10.1037/0003-066X.49.11.929

San Francisco Fire Fighters Local 798 v. San Francisco, 38 Cal. 4th 653 (2006).

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61*, 871–925. doi:10.1111/j.1744-6570.2008.00133.x

Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989).

Washington v. Davis, 426 U.S. 229 (1976).

Watson v. Fort Worth Bank & Trust, 487 U.S. 977 (1988).

Williams et al. v. Ford, 187 F. 3d 533 (1999).